Edge-Cloud Collaborative Multimodal Dehazing for UAV Vision

Junwei Zhao
China Telecom Cloud Computing Research Institute
Beijing, China
zhaojw1@chinatelecom.cn

Qianchun Luo
China Telecom Unmanned Technology Company Ltd.
Nanjing, China
sqlqc@189.cn

Jie Wu†
China Telecom Cloud Computing Research Institute, Beijing, China
Temple University, Philadelphia, USA
jiewu@temple.edu

Abstract

In unmanned aerial vehicle (UAV) applications, dense haze causes strong scattering in the visible (VIS) spectrum, reducing the visibility of small ground-level objects. Infrared (IR) imaging suffers less from haze-induced scattering, but uploading full-frame IR or running large VIS-IR models on board is impractical under tight uplink bandwidth and onboard computational resources. We propose an edge–cloud collaborative framework in which VIS serves as the anchor stream, and only region-of-interest (ROI)-level IR pixels from heavily degraded areas are embedded in-band and transmitted to the cloud for parallel multimodal restoration. The framework comprises three components: (i) edge-side ROI selection via a cross-spectral overlap criterion with geometry-safe anchoring; (ii) single-track in-band ROI delivery via ViStream-SEI with center-out interleaved scheduling; and (iii) cloud-side ROI-parallel enhancement that combines VIS-IR alignment with confidence-aware blending to improve small-object restoration under dense haze. We conduct extensive experiments on multiple public datasets, achieve superior performance, and further validate its effectiveness in real-world scenarios.

CCS Concepts

• **Computing methodologies** → *Knowledge representation and reasoning*; *Computer vision representations.*

Keywords

Edge-Cloud Collaboration; UAV Perception; Multimodal Fusion

ACM Reference Format:
Junwei Zhao, Qianchun Luo, and Jie Wu. 2026. Edge-Cloud Collaborative Multimodal Dehazing for UAV Vision. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), August 9–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3770855.3818393

1 Introduction

Images captured by unmanned aerial vehicles (UAVs) are frequently degraded by dense haze, which is common in high-humidity environments. The resulting scattering in the visible spectrum reduces contrast and suppresses the visibility of small ground-level objects

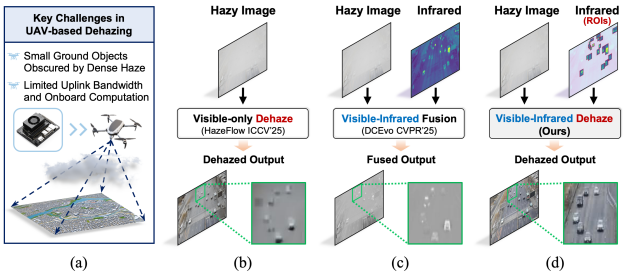


Figure 1: (a) UAV dehazing challenges: dense haze reduces small ground object visibility; uplink bandwidth and onboard computation are limited. (b) Visible-only dehazing [39] struggles in dense haze, leaving residual haze and missing fine structures. (c) Visible-Infrared fusion [26] enhances general features but does not explicitly remove haze, causing appearance distortions. (d) Proposed edge–cloud collaborative method selectively uses ROI-level infrared cues to enhance the visible stream, improving visual quality and small object recovery while remaining computationally efficient.

(Fig. 1a) [15]. This degradation undermines downstream tasks such as surveillance, object detection, and tracking that rely on reliable visual cues [47, 62]. Under clear conditions, visible-light imaging provides rich appearance and detailed textures, but in dense haze it suffers from strong scattering [14, 22]. Thermal infrared imaging is less affected by haze and better preserves coarse structure, yet lacks color and fine textures [3]. These complementary characteristics motivate multimodal fusion of visible and infrared modalities for robust UAV perception in adverse atmospheric conditions.

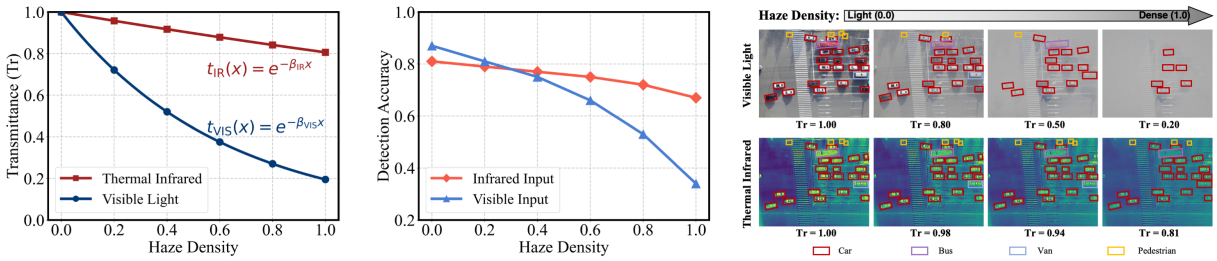
Existing dehazing approaches fall into two categories: visible-only methods and visible-infrared (VIS-IR) fusion methods. (1) Visible-only methods work well in moderate haze, but tend to fail in dense haze, where strong scattering obscures small objects and degrades local context [1, 12] (Fig. 1b). (2) VIS-IR fusion methods mainly focus on feature-level integration and contrast enhancement to facilitate downstream perception tasks such as detection and segmentation [28, 43, 76], rather than explicitly addressing haze removal [24, 59]. Their output often exhibits appearance distortions in dense haze conditions (Fig. 1c). Only a few studies explicitly target multimodal dehazing (e.g. VIFNet [66], HDCFN [75]), yet they do not focus on recovering small ground-level objects, which are particularly challenging in UAV scenarios [42, 63].

In addition, UAV platforms operate under resource constraints: uplink bandwidth is limited with fluctuations, while onboard computing and power budgets are tight [61, 69]. These limits make it difficult to satisfy real-time requirements when running full-capacity dehazing models onboard, while uploading full frame infrared and

† Jie Wu is the corresponding author. This work was supported by the China Telecom Key Research Project on Intelligent Ubiquitous Cloud Technologies (No. 65996).

KDD '26, Jeju Island, Republic of Korea.

ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3818393



(a) Robustness of VIS and IR transmittance to haze (b) Detection accuracy at different haze levels (c) Visualizations of VIS and IR at different haze levels

Figure 2: (a) Visible light undergoes stronger scattering than thermal infrared due to scattering coefficients $\beta_{IR} \ll \beta_{VIS}$. (b) Object detection results: as haze density increases, detection accuracy for visible inputs decreases more rapidly than for infrared. (c) Visualization of infrared and visible imaging effects at the same haze density and their corresponding detection results.

visible streams to the cloud increases latency and could exceed the uplink budget [2]. This motivates an edge–cloud collaborative design that selects informative infrared cues from severely degraded regions and delivers them to the cloud for enhancement.

To address these limitations, we design an edge–cloud collaborative VIS-IR dehazing framework. The visible modality serves as the anchor, and infrared is introduced only in regions the edge device selects as severely hazed. As shown in Fig. 2, infrared wavelengths experience substantially lower scattering than the visible band, and the visibility of small objects decays more slowly as haze increases. The selected infrared regions-of-interest (ROIs) are then carried in-band within a single video stream using supplemental enhancement information (SEI) and scheduled by the center out interleaved scheduling (COIS), which prioritizes coverage of the ROI center and remains effective when the bitstream is truncated by bandwidth limits. On the cloud side, each received IR ROI is processed in parallel with cross spectral alignment and enhancement, followed by confidence-gated aggregation over the full image. These designs reduce uplink load, remain robust to bandwidth fluctuations, and improve small-object restoration under dense haze within the considered computation and bandwidth budgets.

We evaluate dehazing performance on three UAV-captured public datasets (VTUAV, DroneVehicle, and CART). Our method outperforms recent VIS–IR joint dehazing baselines on PSNR, SSIM, and LPIPS, with clearer recovery of small objects under dense haze (Fig. 1d). Under SEI prefix truncation reflecting uplink fluctuations, the proposed COIS improves detection accuracy mAP_{50} by 8.5% at 70% truncation over raster scanning on the VTUAV dataset. On the cloud, ROI-parallel cross-spectral dehazing refines the RGB anchor, and in real foggy urban scenarios the method achieves the lowest NIQE/PIQE (15.70/34.89) among compared methods, with visibly improved small-object clarity.

Our main contributions are summarized as follows:

- We present an edge–cloud collaborative co-design for UAV VIS–IR dehazing under uplink fluctuations, where VIS serves as the anchor stream and IR is offloaded only for selected ROIs to reduce uplink cost under real-time processing constraints.
- We propose ViStream-SEI with COIS, a single-track in-band ROI transport scheme that prioritizes ROI cores and remains robust to SEI prefix truncation under uplink fluctuations.
- We design a VIS–IR differential ROI selection at the edge and pair it with ROI-parallel cross-spectral enhancement on the cloud, to improve small-object restoration under dense haze.
- Extensive experiments on multiple public datasets and real-world UAV scenarios show consistent gains in restoration quality and downstream detection performance, as well as robustness to SEI prefix truncation and VIS–IR misalignment.

2 Related Work

2.1 Image Dehazing

Image dehazing aims to recover clean visual content from haze-degraded input. Recent deep-learning methods have demonstrated strong performance [4, 8, 9, 31, 46, 51, 57]. Representative models include DehazeFormer [41], MB-TaylorFormer [36], SelfPromer [50], PromptHaze [65], and DiffDehaze [55]. Although these models are effective under moderate haze, their performance degrades markedly in dense-haze UAV scenarios, where visible-spectrum signals undergo severe scattering [70]. As a result, they fail to recover fine-scale structures, particularly small objects obscured by haze in aerial imagery. In contrast, we propose a dual-modality framework that leverages infrared cues to enhance the recovery of small-object under severe haze conditions.

2.2 Visible-Infrared Fusion

VIS-IR fusion methods integrate complementary cues to generate improved representations for downstream perception tasks [19, 25, 56, 73]. However, most approaches optimize task features and global contrast rather than performing explicit haze removal [17, 20, 29, 35, 60]. Works that explicitly target VIS-IR dehazing remain at early stage. VIFNet [66] is an early work on feature-level VIS-IR fusion for restoration, with its main focus on general fusion design, while the recovery of fine-scale small objects under dense haze calls for further IR-guided refinement. HDCFN [75] introduces haze-aware adaptive fusion, yet its multi-stage, hierarchical processing and full-frame multimodal inference cause notable computational overhead, making real-time edge-side execution challenging.

2.3 Edge-Cloud Collaboration

Edge–cloud collaboration for vision systems allocates computation between onboard devices and remote servers under bandwidth and latency constraints [7, 30, 40, 52, 54]. Prior studies typically follow three directions: (i) deploying lightweight or compressed models on the edge [11, 21, 27, 37], (ii) offloading full images or features to the cloud [32, 38, 68], and (iii) split computation with task-aware partitioning [10, 53]. Most of these works target instance-level perception tasks such as detection [13, 64, 67], tracking [33], or

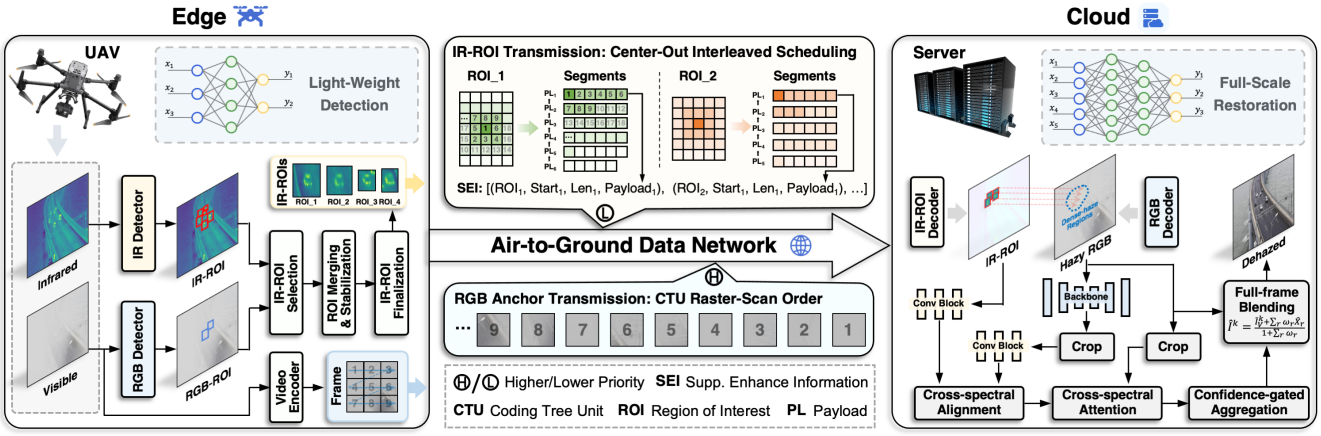


Figure 3: Overview of the proposed edge–cloud collaborative VIS–IR dehazing framework. On the edge, visible and infrared frames are captured, lightweight detectors produce candidate ROIs, and a VIS–IR differential criterion selects IR–ROIs in regions where visible evidence is weak but infrared cues are reliable. Over the air-to-ground link, RGB anchor frames are transmitted together with in-band IR–ROI segments and the inset illustrates center-out, distance-ordered spiral rasterization that prioritizes ROI cores. On the cloud, RGB and IR–ROIs are reassembled and processed in parallel via cross-spectral alignment and attention, followed by confidence-gated aggregation to produce dehazed results.

re-identification [52]. By contrast, pixel-level restoration requires dense predictions with per-pixel intensity, so quality must be preserved in the image domain rather than approximated in the feature space. In addition, VIS–IR fusion requires cross-spectral alignment, which further increases the challenge for edge–cloud collaboration.

3 Methodology

3.1 Overview

We consider UAV–based VIS–IR dehazing. For the k-th frame, the RGB image $I_V^k \in \mathbb{R}^{H\times W\times 3}$ and the IR image $I_{IR}^k \in \mathbb{R}^{H\times W\times 1}$ are temporally synchronized and coarsely registered to $H\times W$. The collaborative system outputs a dehazed RGB image $\hat{I}^k \in \mathbb{R}^{H\times W\times 3}$ under a *real-time* constraint $T_{\text{frame}}^k \le \delta$ (e.g., 33.3 ms) and a *bandwidth* budget. The per-frame bit budget is $C_k = B_k\,\Delta_k$, where B_k is the instantaneous uplink throughput (bit/s) and Δ_k is the available time slot (s). If the RGB anchor consumes D_{rgb}^k bits, the remaining budget for IR-ROI payload is $C_k' = C_k - D_{\text{rgb}}^k$.

To prioritize small-object recovery under joint communication and computation constraints, we decompose the framework (as shown in Fig. 3) into three stages:

(i) ***Edge.*** Given (I_V^k, I_{IR}^k), lightweight detectors together with the VIS–IR confidence–overlap analysis produce a set of differential IR–ROIs,

$$\mathcal{R}_k = \{\, r_t = (\bar{x}_0^t, \bar{y}_0^t, \bar{w}^t, \bar{h}^t)\,\}_{t=1}^{T_k}, \tag{1}$$

where r_t is an axis-aligned rectangle on the $H\times W$ grid with top-left pixel $(\bar{x}_0^t, \bar{y}_0^t)$ and integer width/height $(\bar{w}^t, \bar{h}^t)$. For each IR–ROI, we also output its discrete center c_{r_t} and a persistent identifier roi_id for cross-frame consistency. If $\mathcal{R}_k = \emptyset$, no IR payload is sent.

(ii) ***Edge–Cloud.*** We adopt a single H.26x track: the RGB anchor is the main stream, and IR–ROI pixels are injected in-band as SEI. Let Σ_r denote the center–out spiral index over ROI r induced by its center c_r. A micro–segment (roi_id, start, len, payload) carries len consecutive samples from Σ_r starting at start. For the frame k, the segments and budget satisfy,

$$\mathcal{P}_k = \{(\text{roi\_id, start, len, payload})\},$$
$$L(\mathcal{P}_k) = \sum_{(r,i)\in\mathcal{P}_k} \text{len}_{r,i}\, b_{\text{IR}} + \overline{H}\,|\mathcal{P}_k| \;\le\; C_k', \tag{2}$$

where b_{IR} is the IR bit depth per pixel and $\overline{H}$ is the per–segment header. Center–Out Interleaved Scheduling (COIS) interleaves segments so that the bitstream prefix first covers ROI cores across active ROIs.

(iii) ***Cloud.*** The cloud decodes the RGB anchor $\widetilde{I}_V^k$, reconstructs each IR–ROI patch P_r from SEI, and extracts the corresponding RGB crop $X_r = \widetilde{I}_V^k[\Omega_r]$. It then applies to the following equations,

$$\hat{X}_r = \mathcal{F}_\Theta(X_r, P_r),\;\; \hat{I}^k = \mathcal{B}\big(\widetilde{I}_V^k, \{\hat{X}_r, \Omega_r\}_{r\in\mathcal{R}_k}\big), \tag{3}$$

where $\mathcal{F}_\Theta$ jointly estimates cross-spectral displacement and enhancement, and $\mathcal{B}(\cdot)$ performs confidence-gated, adaptive fusion over the full frame. All ROIs are processed in parallel on GPU, which helps the system meet the real-time constraint $T_{\text{frame}}^k \le \delta$.

Task Formalization. Let $\mathcal{J}(\hat{I}^k; I_{\text{GT}}^k)$ denote the quality functional of dehazing composed of full-frame pixel fidelity, perceptual fidelity, and structural fidelity at ROI-level. The edge-cloud system-level objective can be written as,

$$\min_{\Theta,\,\{\mathcal{P}_k\}} \;\mathbb{E}\big[\mathcal{J}(\hat{I}^k; I_{\text{GT}}^k)\big],$$
$$\text{s.t.}\;\begin{cases} L(\mathcal{P}_k) \le C_k' & \text{(bandwidth budget)},\\ T_{\text{frame}}^k \le \delta & \text{(real-time constraint)},\\ \hat{I}^k = \mathcal{B}\big(\widetilde{I}_V^k, \{\mathcal{F}_\Theta(X_r, P_r)\}_{r\in\mathcal{R}_k}\big) & \text{(parallel fusion)}. \end{cases} \tag{4}$$

3.2 Edge: Differential IR-ROI Generation

At the edge device, under bandwidth and compute budgets, we locate regions where visible evidence is weak but infrared cues are reliable via a VIS–IR differential criterion and, when present, instantiate a set of IR–ROIs $\mathcal{R}_k$. Each ROI is an axis-aligned integer box

r with a discrete center c_r and a persistent roi_id. These variables serve as the sole spatial anchors for subsequent in-band transport and cloud-side processing.

IR–ROI Selection. Given inputs (I_V^k, I_{IR}^k), we apply light-weight object detectors D_V (on I_V^k) and D_{IR} (on I_{IR}^k), finetuned on our VIS and IR datasets. They return candidate sets as follows,

$$\mathcal{B}_V^k = \{(b_j^V, s_j^V)\}_{j=1}^{N_V^k}, \qquad \mathcal{B}_{IR}^k = \{(b_i^{IR}, s_i^{IR})\}_{i=1}^{N_{IR}^k}, \tag{5}$$

where $b = (x_0, y_0, w, h)$ and $s \in [0,1]$ is the detector confidence (set sizes are N_V^k and N_{IR}^k). To reduce inter–scene scale variation, we apply per–frame min–max normalization to obtain $\hat{s}_j^V, \hat{s}_i^{IR} \in [0,1]$. Let the spatial overlap between two boxes be measured by Intersection-over-Union (IoU) $\mathrm{IoU}(b, b') = |b \cap b'|/|b \cup b'|$. For each b_i^{IR}, we quantify its maximum spatial agreement with b_j^V as,

$$m_i^k = \max_{(b_j^V, \hat{s}_j^V) \in \mathcal{B}_V^k} \mathrm{IoU}\big(b_i^{IR},\, b_j^V\big). \tag{6}$$

We define the cross–spectral differential score as,

$$\Delta_i^k = \hat{s}_i^{IR}\,\big(1 - m_i^k\big) \in [0,1], \tag{7}$$

which increases when an IR detection is confident yet lacks a good VIS counterpart. IR candidate is retained if,

$$z_i^k = \begin{cases} 1, & \Delta_i^k \geq \theta_{\mathrm{diff}} \ \wedge\ |b_i^{IR}| \geq A_{\min}, \\ 0, & \text{otherwise}, \end{cases} \tag{8}$$

where threshold $\theta_{\mathrm{diff}} \in (0,1)$, $|b| = wh$ denotes the box area, and $A_{\min}$ is the minimum area for a valid IR candidate.

ROI Merging and Stabilization. On the retained IR candidates $\{b_i^{IR} \mid z_i^k = 1\}$, we first apply standard non-maximum suppression (NMS) in descending Δ_i^k with an IoU threshold of 0.5, which is used as an empirical setting in detection. We then perform connected merging: boxes whose pairwise IoU exceeds 0.5 are grouped and each group is replaced by the tight rectangle of its union, yielding merged ROIs $r_t = (x_0^t, y_0^t, w^t, h^t)$. To accommodate small registration errors and to align with codec block boundaries, each r_t is isotropically dilated by 20% as,

$$\begin{aligned} \bar{w}^t &= \min\Big(W,\ \big\lceil 1.2\, w^t \big\rceil\Big),\ \ \bar{h}^t = \min\Big(H,\ \big\lceil 1.2\, h^t \big\rceil\Big), \\ \bar{x}_0^t &= \mathrm{snap}\Big(x_0^t - \Big\lfloor \tfrac{\bar{w}^t - w^t}{2} \Big\rfloor\Big),\ \ \bar{y}_0^t = \mathrm{snap}\Big(y_0^t - \Big\lfloor \tfrac{\bar{h}^t - h^t}{2} \Big\rfloor\Big), \end{aligned} \tag{9}$$

where $\mathrm{snap}(u) = 16 \cdot \lfloor u/16 \rfloor$ aligns the coordinates with the codec's block grid.

IR–ROI Finalization. The stabilized set of IR–ROIs is defined as,

$$\mathcal{R}_k = \big\{\, (\bar{x}_0^t, \bar{y}_0^t, \bar{w}^t, \bar{h}^t) \;\big|\; t = 1, \ldots, T_k \,\big\}, \tag{10}$$

and each ROI center is $c_{r_t} = \Big(\Big\lfloor \bar{x}_0 + \frac{\bar{w}-1}{2} \Big\rfloor, \Big\lfloor \bar{y}_0 + \frac{\bar{h}-1}{2} \Big\rfloor\Big)$. ROI IDs roi_id persist across frames via one-to-one IoU association. Unmatched current ROIs are assigned new IDs. Algorithm 1 summarizes the edge-side IR–ROI generation procedure.

Algorithm 1 Edge-side Differential IR-ROI Generation

Require: Registered (I_V^k, I_{IR}^k); object detectors D_V, D_{IR}; threshold θ_{diff}; minimum area $A_{\min}$
Ensure: IR–ROI set $\mathcal{R}_k$, centers $\{c_r\}$, and roi_id

$\mathcal{B}_V^k \leftarrow D_V(I_V^k)$, $\mathcal{B}_{IR}^k \leftarrow D_{IR}(I_{IR}^k)$ (candidates per (5))
frame-wise normalize scores to get $\{\hat{s}_j^V\}, \{\hat{s}_i^{IR}\}$
for each $(b_i^{IR}, \hat{s}_i^{IR}) \in \mathcal{B}_{IR}^k$ **do**
    compute m_i^k by (6); compute Δ_i^k by (7)
    set z_i^k by trigger rule (8)
$C \leftarrow \{\, b_i^{IR} \mid z_i^k = 1 \,\}$
apply standard NMS on C and sorted by Δ_i^k descending
geometry stabilization per (9) to get $\mathcal{R}_k$; compute c_r
associate $\mathcal{R}_k$ with $\mathcal{R}_{k-1}$ by max-IoU; assign roi_id
return $\mathcal{R}_k$, $\{c_r\}$, roi_id

3.3 Edge-Cloud: ViStream-SEI (SEI-Based Single Track ROI Transport with COIS)

We design *ViStream-SEI*, a single-track H.26x transport scheme in which the full-frame RGB anchor is encoded as the primary stream, while IR-ROI pixels are packed into SEI units. Under the residual uplink budget C_k', each ROI is serialized into spiral-ordered micro-segments and interleaved via COIS, so that any bitstream prefix prioritizes ROI cores (schematized in Fig. 3, middle).

Center-Out Interleaved Scheduling. Given a stabilized ROI $r = (x_0^r, y_0^r, w^r, h^r)$ (bars in Eq. (10) omitted hereafter for brevity), let the ROI-local index set be $\Omega_r = \{(u,v) :\ 0 \leq u < w^r,\ 0 \leq v < h^r\}$ and the local center $\tilde{c}_r = (c_x^r - x_0^r,\ c_y^r - y_0^r)$. The Chebyshev radius and the clockwise shells are given by,

$$\begin{aligned} \rho_r(u,v) &= \max\{|u - \tilde{c}_x^r|,\ |v - \tilde{c}_y^r|\}, \\ \mathcal{S}_\ell(r) &= \{(u,v) \in \Omega_r : \rho_r(u,v) = \ell\} \text{ (clockwise)}, \\ \Sigma_r &= \mathrm{Concat}\big(\mathcal{S}_0(r), \mathcal{S}_1(r), \ldots\big), \end{aligned} \tag{11}$$

where Σ_r is a bijection from Ω_r to $\{0, ..., N_r - 1\}$ with $N_r = w^r h^r$. The edge partitions Σ_r into micro-segments and COIS interleaves active-ROI segments by ascending shell index (core-first), breaking ties round-robin.

ROI States and Geometry Signaling. Each ROI carries a state $\in$ {new, update, done, idle}. Geometry $(x_0^r, y_0^r, w^r, h^r, c_x^r, c_y^r)$ is signaled only in new and update states. In state done, the edge additionally sets a 1-bit completion flag fin_r in SEI, indicating that all N_r pixels of r are included in the current frame. Pixel micro-segments are carried in new, update, and done, while idle transmits no payload. The cloud maintains a geometry map keyed by roi_id and marks an ROI complete when its buffer accumulates N_r samples or when the completion flag is received.

SEI Encapsulation and Parsing. Pixel micro-segments and, when needed, geometry messages are carried as user_data_unregistered SEI NAL units in the same video track and may be split across multiple SEI NALs if required. Upon reception, the cloud parses SEI and, using (roi_id, start, len, payload) together with the geometry map, places samples into per-ROI buffers to reconstruct P_r.

Table 1: Quantitative comparison results of dense haze removal and complexity on UAV datasets. Higher PSNR/SSIM and lower LPIPS indicate better quality. Best results are bolded. (Edge runtime N/A when the model exceeds Orin Nano's unified memory.)

Method	Reference	Publication	Year	Modality	VTUAV			DroneV			CART			Complexity			
					PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	Params	GFLOPs	Edge-t	Cloud-t
SFSNiD	Cong et al. [5]	CVPR	2024	VIS	26.18	0.864	0.382	18.53	0.665	0.561	20.85	0.801	0.435	8.35M	557.4	N/A	219ms
ConvIR	Cui et al. [6]	T-PAMI	2024	VIS	28.53	0.872	0.349	18.74	0.672	0.534	21.07	0.799	0.413	8.63M	374.5	785ms	94ms
DFG-DDM	Li et al. [18]	T-GRS	2025	VIS	28.64	0.871	0.342	18.93	0.674	0.531	21.12	0.802	0.409	112.9M	2161.8	N/A	613ms
HazeFlow	Shin et al. [39]	ICCV	2025	VIS	28.80	0.876	0.335	19.18	0.677	0.529	21.47	0.807	0.396	50.86M	915.3	N/A	274ms
VIFNet	Yu et al. [66]	Neurocomp.	2024	VIS+IR	29.03	0.879	0.297	20.51	0.703	0.432	22.38	0.825	0.344	9.78M	708.3	N/A	152ms
IASSF	Zhang et al. [74]	PR	2026	VIS+IR	29.21	0.882	0.285	20.74	0.706	0.428	22.57	0.833	0.330	9.21M	518.4	237ms	115ms
HDCFN	Zhao et al. [75]	ACM MM	2025	VIS+IR	30.06	0.887	0.274	21.06	0.719	0.416	22.85	0.847	0.325	6.81M	305.1	128ms	63ms
Ours	–	–	2026	VIS+IR	**31.92**	**0.904**	**0.238**	**22.84**	**0.743**	**0.397**	**23.19**	**0.862**	**0.316**	**3.47M**	**273.6**	**26ms**	**31ms**

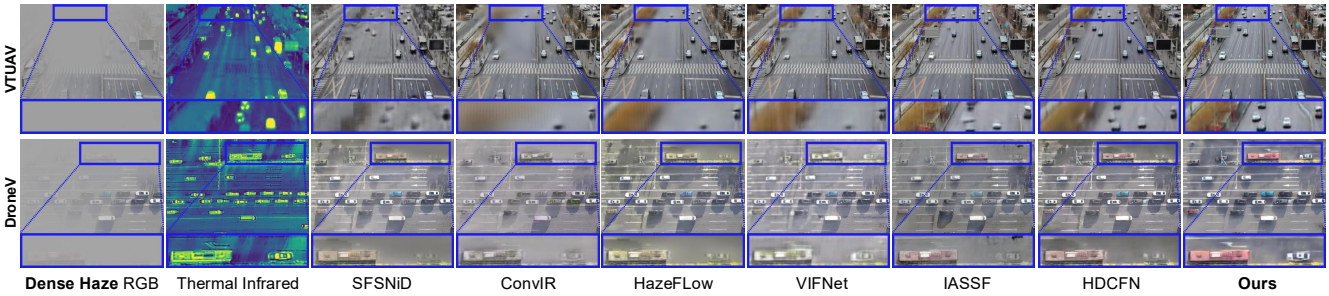


Figure 4: Visualization of representative dehazing results on VTUAV and DroneV datasets.

3.4 Cloud: ROI-Parallel Cross-modal Alignment and Enhancement

Decoding yields the RGB anchor $\widetilde{I}_V^k \in \mathbb{R}^{H\times W\times 3}$ and the ROI set $\mathcal{R}_k = \{r\}$. For each $r \in \mathcal{R}_k$, we obtain the RGB crop $X_r = \widetilde{I}_V^k[\Omega_r] \in \mathbb{R}^{h_r\times w_r\times 3}$ and the IR patch $P_r \in \mathbb{R}^{h_r\times w_r\times 1}$. The objective is to produce a dehazed RGB output $\hat{I}^k$, with IR-augmented refinement concentrated inside ROIs.

Cross-spectral Alignment and Attention. We compute a lightweight two-level RGB pyramid once per frame as follows,

$$\mathcal{G}^k = \Psi_V(\widetilde{I}_V^k) = \{G^{(0)}, G^{(1)}\},\ G^{(1)} = \mathrm{Down}_2(G^{(0)}), \tag{12}$$

where Ψ_V denotes a lightweight backbone built from depthwise-separable convolutions, and Down_2 is a stride-2 downsampling operation. For each ROI r, we crop ROI features from each pyramid level $G^{(s)} \in \mathcal{G}^k$ as $G_r^{(s)} = \mathcal{C}(G^{(s)}, \Omega_r)$, $s \in \{0, 1\}$.

The isotropic dilation in Eq. (9) provides a geometry-safe margin. We choose radii $(\rho^{\downarrow}, \rho^{\uparrow})$ such that $\rho^{\downarrow} + \rho^{\uparrow}$ does not exceed this margin. At the coarse level,

$$C_r^{(1)}(x, y, \Delta u, \Delta v) = \left\langle \phi(G_r^{(1)}(x, y)), \psi(P_r^{(1)}(x + \Delta u, y + \Delta v)) \right\rangle, \tag{13}$$

where $|\Delta u|, |\Delta v| \le \rho^{\downarrow}$, with channel-wise dot product $\langle\cdot,\cdot\rangle$ and $P_r^{(1)} = \mathrm{Down}_2(P_r)$. Both VIS and IR are projected into a shared embedding via non-shared heads ϕ, ψ (conv blocks). A lightweight 3D convolutional head regresses coarse displacement/confidence $(\Phi_r^{(1)}, C_r^{(1)})$. At full resolution, we refine around $(\Phi_r^{(1)}\uparrow)$ within $|\delta u|, |\delta v| \le \rho^{\uparrow}$ to obtain $(\delta\Phi_r, C_r^{(0)})$, and then fuse the coarse and fine confidences with 1×1 convolution f_{fuse} followed by sigmoid σ,

$$\Phi_r = \Phi_r^{(1)}\uparrow + \delta\Phi_r,\ C_r = \sigma\Big(f_{\text{fuse}}\big([C_r^{(0)},\ C_r^{(1)}\uparrow]\big)\Big), \tag{14}$$

where out-of-bound samples use reflection padding. We then warp the IR patch P_r by the displacement field Φ_r via $\mathcal{W}$ as follows,

$$\widetilde{P}_r = \mathcal{W}(P_r, \Phi_r), \tag{15}$$

and apply a single-head, depthwise-separable cross-spectral attention with the learnable offset $\Delta = \mathrm{Conv}_{3\times3}([\Phi_r, \nabla X_r, \nabla\widetilde{P}_r])$ (where ∇ is the Sobel operator). Let $Q = W_Q * X_r$, $K = W_K * \widetilde{P}_r$, $V = W_V * \widetilde{P}_r$. Offsets are clipped to a fixed window before sampling. The residual is injected as,

$$\hat{X}_r = X_r + M_r \odot W_O(\mathcal{A} \odot V),\ \mathcal{A}(x) = \mathbb{S}\Big(\frac{Q(x)^\top K(x+\Delta(x))}{\sqrt{d_h}}\Big), \tag{16}$$

where $\mathbb{S}$ denotes the Softmax used for normalization, $*$ denotes convolution, d_h is the head width and $\odot$ is element-wise product.

Confidence-gated Aggregation. We decouple two roles: M_r controls how much IR residual is injected inside each ROI, and the blending rule sets how ROI results mix with the RGB anchor over full frame. Let $\tilde{d}_b = d_{\partial r}/R_b$ and $\tilde{d}_c = d_{c_r}/R_c$. The per-pixel gate is,

$$\bar{C}_r = \sigma\Big(\tfrac{C_r - \mathrm{mean}(C_r)}{\mathrm{std}(C_r)+10^{-6}}\Big),\quad M_r = \bar{C}_r\,\kappa_b\,\kappa_c, \tag{17}$$

where $\kappa_b = \cos^2\big(\frac{\pi}{2}\tilde{d}_b\big)$ and $\kappa_c = \cos^2\big(\frac{\pi}{2}\tilde{d}_c\big)$, $R_b = \frac{1}{2}\min(h_r, w_r)$ and $R_c = \frac{1}{2}\max(h_r, w_r)$. Here $\bar{C}_r$ is a self-normalized confidence in $[0, 1]$ (z-score followed by a sigmoid). κ_b softly attenuates near ROI boundaries to avoid seams. κ_c emphasizes ROI cores where alignment is typically most reliable.

Full-frame blending uses confidence weights as,

$$\hat{I}^k(p) = \frac{\widetilde{I}_V^k(p) + \sum_{r\ni p}\omega_r(p)\,\hat{X}_r(p)}{1 + \sum_{r\ni p}\omega_r(p)},\ \omega_r(p) = \exp\big(\bar{C}_r(p)\big) - 1, \tag{18}$$

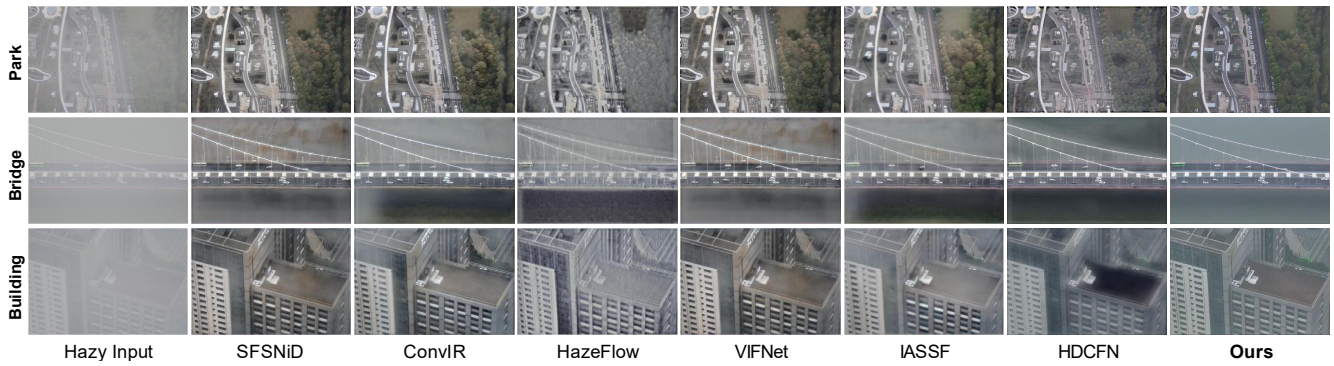


Figure 5: Visualization of representative dehazing results in real-world urban scenes under foggy conditions captured by a UAV.

where $\forall p \in \Omega$. This softmax-like weighting prioritizes high-confidence ROIs, while the unit term in the denominator provides a stable fallback to the RGB anchor. Overlaps are naturally normalized to avoid over-amplification. If no ROI covers p, then $\hat{I}^k(p) = \widetilde{I}_V^k(p)$.

Training Objective. We optimize three complementary terms, i.e., pixel fidelity, perceptual fidelity, and ROI-structural fidelity. To reduce the contribution of non-updatable regions, we apply a mild ROI up-weighting $\omega(p) = 1 + \mathbf{1}[p \in \cup_r \Omega_r]$ and normalize by the total weight.

Pixel fidelity term $\mathcal{L}_{\text{pix}}$ is defined as,

$$\mathcal{L}_{\text{pix}} = \frac{1}{\sum_{p\in\Omega} \omega(p)} \sum_{p\in\Omega} \omega(p)\, \rho\big(\hat{I}^k(p) - I_{\text{GT}}^k(p)\big), \tag{19}$$

where $\rho(z) = \sqrt{z^2 + 10^{-6}}$, Ω is the full-frame pixel set.

Perceptual fidelity term $\mathcal{L}_{\text{perc}}$ is defined as,

$$\mathcal{L}_{\text{perc}} = \sum_{\ell\in\mathcal{K}} \frac{1}{C_\ell H_\ell W_\ell} \left\| W_\ell \odot \big(\Phi_\ell(\hat{I}^k) - \Phi_\ell(I_{\text{GT}}^k)\big)\right\|_2^2, \tag{20}$$

where $\mathcal{K}$ is the set of selected layers of a frozen feature extractor $\Phi_\ell(\cdot)$, and W_ℓ is the downsampled ROI weight.

ROI structural fidelity term $\mathcal{L}_{\text{roi}}$ is defined as,

$$\mathcal{L}_{\text{roi}} = \frac{1}{\sum_{r\in\mathcal{R}_k} |\Omega_r|} \sum_{r\in\mathcal{R}_k} \sum_{x\in\Omega_r} \left\| \nabla\hat{I}^k(x) - \nabla I_{\text{GT}}^k(x)\right\|_1, \tag{21}$$

where ∇ is the Sobel operator and $\mathcal{L}_{\text{roi}}$ is defined as zero for $\mathcal{R}_k = \emptyset$.

The total loss $\mathcal{L}_{\text{cloud}}$ can be formulated as,

$$\mathcal{L}_{\text{cloud}} = \mathcal{L}_{\text{pix}} + \lambda_{\text{perc}}\, \mathcal{L}_{\text{perc}} + \lambda_{\text{roi}}\, \mathcal{L}_{\text{roi}}. \tag{22}$$

Following common practice, we set $\lambda_{\text{perc}} = 0.05$ and $\lambda_{\text{roi}} = 0.5$, and observe stable performance for $\lambda_{\text{perc}} \in [0.03, 0.07]$ and $\lambda_{\text{roi}} \in [0.2, 0.8]$ in a small sweep of sensitivity.

4 Experiment

4.1 Experimental Settings

We conduct experiments on three publicly available, real-world, synchronized VIS–IR datasets, all captured using UAVs. Haze synthesis is implemented based on a depth-assisted atmospheric scattering model [22]. An overview of datasets is as follows:

Public Datasets. (i) *VTUAV* [71] dataset provides paired visible and infrared UAV images across diverse environments, including urban and rural scenes with cars, parks, pedestrians, and streets. (ii) *DroneVehicle* [44] dataset (abbreviated as *DroneV*) contains VIS-IR image pairs captured by drone-mounted cameras, featuring various vehicle types (e.g. cars, trucks, busses, vans) under different lighting conditions. (iii) *CART* [16] dataset offers high-resolution paired VIS-IR images captured in natural environments, covering a diverse range of terrains such as rivers, bridges, rocks and lakes.

Table 2: Evaluation results on real-world foggy urban scenes.

Method	SFSNiD	ConvIR	HazeFlow	VIFNet	IASSF	HDCFN	**Ours**
NIQE ↓	19.40±3.1	18.83±3.0	18.42±2.9	16.94±2.6	16.37±2.7	16.15±2.4	**15.70**±2.3
PIQE ↓	40.62±8.2	38.55±7.8	37.73±7.3	37.11±6.9	36.84±7.1	36.21±6.5	**34.89**±6.2

Implementation. The proposed method is implemented using PyTorch. For runtime profiling, we measure the cloud side on an NVIDIA A100 40G GPU and the edge side on an NVIDIA Jetson Orin Nano. The models are optimized using the Adam optimizer ($\beta_1 = 0.9$ and $\beta_2 = 0.999$). The initial learning rate is set to 1e-4 and gradually reduced to 1e-6 using cosine annealing. Batch size is configured to 4 during training. The input training images are resized to 512×512.

4.2 Comparison with Recent Methods

We conducted comparisons with seven recent methods: SFSNiD [5], ConvIR [6], DFG-DDM [18], HazeFlow [39], VIFNet [66], IASSF [74] and HDCFN [75]. All methods were retrained on VTUAV, DroneV and CART using consistent settings, and hyperparameters were adopted from official implementations. We evaluated performance using PSNR [58], SSIM [58] and LPIPS [72] metrics. Qualitative and quantitative results are shown in Fig. 4 and Tab. 1, respectively.

(i) *Visual Quality.* Results in Tab. 1 show that our method achieves superior dehazing across three UAV datasets, yielding higher PSNR and SSIM and lower LPIPS under dense haze. Qualitative examples in Fig. 4 exhibit clearer boundaries and better preservation of fine structures, especially in severely occluded regions.

(ii) *Complexity.* We report model parameters (Params), computation (FLOPs), and runtime (Edge-t/Cloud-t) per frame (512×512, batch = 1) on Jetson Orin Nano (edge) and NVIDIA A100 (cloud). As summarized in Tab. 1, the collaborative design attains the lowest latency among compared methods and meets real-time ($\lesssim$33.3 ms) on the edge platform. Under pipelined execution between edge and

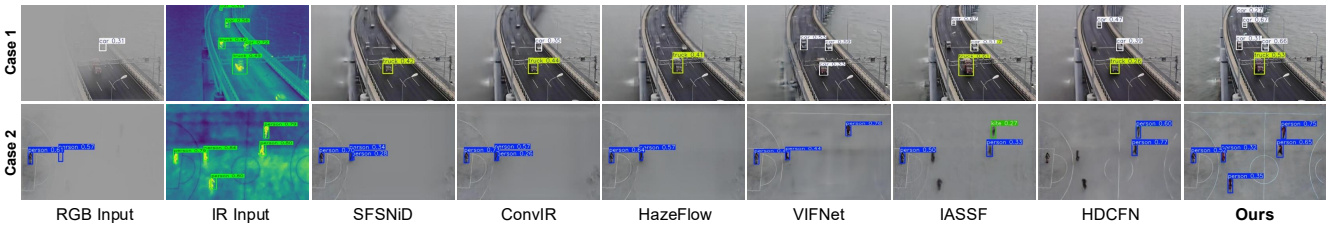


Figure 6: Visualization of object detection results on the outputs generated by dehazing methods.

Table 3: Evaluation results on object detection task.

Method	SFSNiD	ConvIR	HazeFlow	VIFNet	IASSF	HDCFN	**Ours**
mAP_{50}	0.512	0.528	0.541	0.564	0.570	0.581	**0.632**
mAR_{50}	0.586	0.603	0.612	0.645	0.653	0.656	**0.702**

Table 4: Effectiveness of IR-ROIs on dehazing results.

Case	PSNR	SSIM	LPIPS	mAP_{50}	mAR_{50}	runtime
VIS-only	28.63	0.859	0.384	0.527	0.604	24ms
+ IR-full	32.07	0.911	0.234	0.635	0.709	52ms
+ IR-ROIs	31.92 -0.15	0.904 -0.007	0.238 +0.004	0.632 -0.003	0.702 -0.007	31ms -21ms

cloud, the throughput is governed by max(Edge-t, Cloud-t) rather than their sum. Efficiency comes from lightweight ROI detection on the edge and parallel ROI processing on the cloud, yielding a stronger speed–quality trade-off than single-device baselines.

4.3 Evaluation on Real-World Scenarios

For real-world evaluation, we use a Zenmuse H20T-equipped UAV operating across diverse foggy urban scenes, covering approximately 15,000 paired VIS–IR frames in our comparative experiments. We follow no-reference IQA practice and report NIQE [34] and PIQE [49] (lower is better) for quantitative comparison.

Tab. 2 reports quantitative results, and Fig. 5 provides visual comparisons. Our method achieves the best NIQE/PIQE (↓), e.g., 15.70 vs. 16.15 and 34.89 vs. 36.21 against the second-best method (HDCFN), indicating our outputs closer to natural image statistics with fewer perceptual artifacts. Qualitatively, our method preserves richer fine textures (e.g., small cars) and sharper edges while maintaining structural consistency and realistic color fidelity, suggesting better generalization to real-world UAV scenes.

4.4 Evaluation on Object Detection Task

To assess whether dehazing-induced visibility improvements contribute to downstream perception, we evaluate object detection performance on VTUAV. Following prior protocols (e.g., CORUN [8], DIVFusion [45]), we use a standard YOLOv8m detector [48] pretrained on COCO [23] and fine-tuned on VTUAV.

Fig. 6 shows the detection results on dehazed frames. Compared methods retain haze or oversmooth in heavily occluded regions, producing uncertain edges. Ours enhances local contrast and preserves structure, providing more reliable cues for detectors. Quantitatively (Tab. 3), we achieve the highest mAP_{50} and mAR_{50}. The higher value of mAR_{50} indicates fewer missed small or low-contrast targets, reflecting improved recall. These outcomes align with our design, which integrates edge-side IR-ROI selection and cloud-side cross-spectral alignment with confidence-gated blending, supporting higher detection accuracy under dense haze.

Table 5: ROI params sweep: θ_{diff} and $A_{\min}$ (SSIM/mAP_{50}).

$A_{\min}$	$\theta_{\text{diff}}=0.2$	$\theta_{\text{diff}}=0.4$	$\theta_{\text{diff}}=0.6$	$\theta_{\text{diff}}=0.8$
256 px²	0.901/0.626	0.902/0.629	0.898/0.619	0.891/0.606
576 px²	0.902/0.628	**0.904**/**0.632**	0.900/0.625	0.893/0.610
1024 px²	0.895/0.618	0.897/0.623	0.891/0.611	0.885/0.602

Table 6: The COIS compared to other scheduling strategies.

Scan (Strategy)	70% trunc.		40% trunc.		10% trunc.	
	SSIM	mAP_{50}	SSIM	mAP_{50}	SSIM	mAP_{50}
Raster	0.862	0.538	0.876	0.574	0.890	0.607
Zig–Zag	0.867	0.546	0.881	0.582	0.895	0.612
Serpentine	0.871	0.552	0.885	0.588	0.898	0.616
COIS (ours)	**0.879**	**0.584**	**0.894**	**0.609**	**0.901**	**0.625**

4.5 Ablation Studies

We conduct ablations over the edge-side ROI selection, edge–cloud transport, cloud-side modules, and loss functions to verify the contribution of each component.

(i) *Edge-side ROI Selection.* Fig. 2 shows wavelength-dependent transmittance. IR is less attenuated than VIS under heavy haze, yielding more stable detections of small targets. As in Tab. 4, using IR-full-frame attains the highest scores with 52 ms cloud time. IR–ROIs achieve competitive results while significantly reducing cloud time to 31 ms. This indicates that informative IR evidence could be spatially sparse and effectively captured by the IR-ROI selection. Tab. 5 shows an optimum at $A_{\min} = 576$ px², $\theta_{\text{diff}} = 0.4$. Smaller $A_{\min}$ (256) underperforms due to fragmented ROIs and noisier alignment. Larger $A_{\min}$ (1024) misses small targets and degrades SSIM and mAP_{50}. Increasing θ_{diff} from 0.4 to 0.8 reduces trigger recall. Overall, performance is stable and near-optimal in the neighborhood of $A_{\min}=576$ with $\theta_{\text{diff}}=0.4$.

(ii) *Edge-Cloud Transport.* To assess robustness under bandwidth truncation, we simulate IR–SEI prefix truncation at 10%, 40%, and 70% on VTUAV and compare Raster, Zig–Zag, Serpentine, and COIS (Tab. 6). COIS yields the highest SSIM and mAP_{50} at all truncation levels, with the largest gains at 70%: +0.017 SSIM and +0.046 mAP_{50} (+8.5%) over Raster. Gains taper as the prefix grows (40%: +0.012 SSIM / +3.6% mAP_{50}; 10%: +0.006 SSIM / +1.6% mAP_{50}), consistent with diminishing benefit once most ROI cores are received. The advantage stems from COIS's center-out interleaving across ROIs, which delivers core pixels early so short prefixes already carry semantically useful content under bandwidth fluctuation.

Table 7: Effectiveness of cloud-side algorithmic modules.

Case	Modules			VTUAV		DroneV	
	Align.	Atten.	Aggre.	PSNR	SSIM	PSNR	SSIM
VIS-only				28.63	0.859	18.81	0.674
(A)	✓			29.55	0.874	19.98	0.695
(B)	✓	✓		31.73	0.899	22.56	0.738
(C)	✓	✓	✓	**31.92**	**0.904**	**22.84**	**0.743**

Table 8: Effectiveness of loss terms on dehazing results.

Case	VTUAV			DroneV		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
$\mathcal{L}_{\text{pix}}$	31.75	0.898	0.267	22.68	0.734	0.426
$+\mathcal{L}_{\text{perc}}$	31.79	0.901	0.243	22.69	0.738	0.404
$+\mathcal{L}_{\text{perc}}+\mathcal{L}_{\text{roi}}$	**31.92**	**0.904**	**0.238**	**22.84**	**0.743**	**0.397**

Table 9: Robustness to VIS-IR misalignment with Temporal offsets (T) of $\pm 1/2$ frames and Spatial shifts (S) of 5/10/15 px.

Method	Registered		T (±1 f)		T (±2 f)		S (5 px)		S (10 px)		S (15 px)	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
VIFNet	29.03	0.879	28.71	0.874	28.26	0.869	28.10	0.865	27.72	0.853	27.44	0.839
IASSF	29.21	0.882	28.97	0.877	28.63	0.872	28.45	0.868	27.89	0.857	27.68	0.851
HDCFN	30.06	0.887	29.53	0.882	28.95	0.874	28.81	0.870	28.03	0.859	27.76	0.854
Ours	**31.92**	**0.904**	**31.70**	**0.902**	**31.35**	**0.899**	**31.04**	**0.896**	**30.29**	**0.884**	**29.31**	**0.870**

(iii) *Cloud-side Modules.* Tab. 7 evaluates Cross-spectral Alignment (Align.), Attention (Atten.), and confidence-gated Aggregation (Aggre.). Align. provides the reliable base gain under VIS–IR misregistration. Compared with VIS-only, it yields +0.92 dB PSNR and +0.015 SSIM on VTUAV. Adding Atten. brings improvement by injecting well-aligned IR residuals. Aggre. further stabilizes overlaps and suppresses artifacts, giving a consistent final lift. Combined, the three modules achieve the superior perceptual quality and stable reconstruction at ROI boundaries.

(iv) *Training Objective.* Tab. 8 reports ablation results for the loss design. Adding $\mathcal{L}_{\text{perc}}$ to $\mathcal{L}_{\text{pix}}$ improves VTUAV by +0.04 dB PSNR, +0.003 SSIM, and −0.024 LPIPS (similar trend on DroneV), indicating sharper textures. Adding $\mathcal{L}_{\text{roi}}$ further enhances ROI edges and yields additional gains on VTUAV (+0.15 dB PSNR, +0.005 SSIM, −0.007 LPIPS). Overall, the combined objective $\mathcal{L}_{\text{pix}}+\mathcal{L}_{\text{perc}}+\mathcal{L}_{\text{roi}}$ provides a robust balance of pixel fidelity, perceptual quality, and ROI detail across datasets.

(v) *Robustness to temporal/spatial misalignment.* In Tab. 9, our method outperforms strong VIS–IR baselines under the same perturbations. This robustness is enabled by confidence-aware alignment and gated IR injection with anchor fallback, which down-weights unreliable IR cues as misalignment increases.

4.6 Uplink Payload and Transmission Time

To quantify the bandwidth cost of different collaboration strategies, we analyze the uplink payload and transmission time for 512×512 inputs under a simplified model, as summarized in Table 10 and shown in Fig. 7. We consider three representative uplink rates $R_{\text{up}} \in \{1, 2, 3\}$ Mbps, covering a low-rate case (1 Mbps), a moderate regime around 2 Mbps, and a higher-rate case around 3 Mbps that is feasible under favorable wireless conditions [77]. For comparability, the RGB anchor stream is fixed at 0.8 Mbps (26.7 kbit per frame at 30 fps) for all methods. Table 10 reports the additional IR payload

Table 10: Illustrative uplink payloads and transmission time per frame under different uplink rates $R_{\text{up}} \in \{1, 2, 3\}$ Mbps. B_{RGB} denotes the per-frame RGB anchor payload, and B_{IR} denotes the additional IR payload.

Method	R_{up} (Mbps)	B_{RGB} (kbit/fr.)	B_{IR} (kbit/fr.)	T_{IR} (ms/fr.)	T_{Total} (ms/fr.)
VIS-only	1	26.7	0.0	0.0	26.7
VIS+IR-full	1	26.7	65.5	65.5	92.2
VIS+IR-ROI	1	26.7	6.6	6.6	33.3
VIS-only	2	26.7	0.0	0.0	13.3
VIS+IR-full	2	26.7	65.5	32.8	46.1
VIS+IR-ROI	2	26.7	6.6	3.3	16.6
VIS-only	3	26.7	0.0	0.0	8.9
VIS+IR-full	3	26.7	65.5	21.8	30.7
VIS+IR-ROI	3	26.7	6.6	2.2	11.1

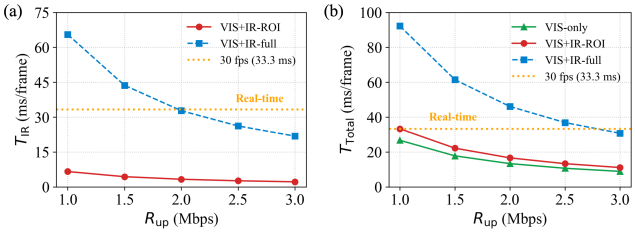


Figure 7: Uplink transmission time for different VIS–IR configurations on 512×512 VTUAV inputs.

B_{IR} together with its uplink transmission time $T_{\text{IR}} = B_{\text{IR}}/R_{\text{up}}$ and the total transmission time $T_{\text{Total}} = (B_{\text{RGB}} + B_{\text{IR}})/R_{\text{up}}$.

The VIS-only configuration incurs no IR uplink cost ($B_{\text{IR}} = 0$), so T_{Total} is determined by the RGB anchor stream. In the VIS+IR-full baseline, full-frame IR requires $B_{\text{IR}} = 65.5$ kbit per frame, leading to substantial uplink overhead, as shown in blue curve in Fig. 7 (a,b). At low bandwidth, uplink transmission exceeds the 33.3 ms frame interval for 30 fps, leaving limited budget for edge/cloud inference.

In contrast, the proposed VIS+IR-ROI configuration concentrates IR bits on sparse task-relevant regions, yielding a much smaller IR payload of 6.6 kbit per frame under the same coding setting. Consequently, T_{IR} is only 6.6, 3.3, and 2.2 ms at $R_{\text{up}} = 1, 2, 3$ Mbps, and the corresponding total transmission times T_{Total} are 33.3, 16.6, and 11.1 ms (Table 10; red curve in Fig. 7 (a,b)). Fig. 7 (b) further shows that the ROI-based curve remains close to the VIS-only curve while staying well below the 33.3 ms reference line at moderate and high bandwidths. Overall, ROI-based VIS–IR collaboration substantially reduces IR uplink time compared to full-frame IR streaming, while keeping the total per-frame transmission budget within a regime compatible with real-time UAV operation.

5 Conclusion

We present an edge-cloud collaborative framework for UAV dehazing that uses the VIS stream as anchor and transmits IR cues only from severely degraded regions. ViStream-SEI embeds IR ROI pixels in-band within a single H.26x stream using COIS for truncation-robust delivery. The cloud performs ROI-parallel alignment and confidence-gated aggregation. Across multiple public datasets and real-world UAV scenarios, our method improves dehazing quality and small-object detection, remains robust to bandwidth truncation, and sustains real-time throughput.

References

[1] Codruta Ancuti, Cosmin Ancuti, Mateu Sbert, and Radu Timofte. 2019. Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images. In *IEEE International Conference on Image Processing*. 1014–1018.

[2] Quan Chen, Ming Yi, Jing Li, Ning Li, Hong Gao, and Zhipeng Cai. 2025. Optimal and Approximate Parallelism-Based Computation Offloading Algorithms for Real-Time Multimodal Learning at the Edge. In *IEEE INFOCOM 2025-IEEE Conference on Computer Communications*. IEEE, 1–10.

[3] Tianxiang Chen, Qi Chu, Bin Liu, and Nenghai Yu. 2023. Fluid Dynamics-Inspired Network for Infrared Small Target Detection.. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 590–598.

[4] Zixuan Chen, Zewei He, and Zhe-Ming Lu. 2024. DEA-Net: Single image dehazing based on detail-enhanced convolution and content-guided attention. *IEEE Transactions on Image Processing* 33 (2024), 1002–1015.

[5] Xiaofeng Cong, Jie Gui, Jing Zhang, Junming Hou, and Hao Shen. 2024. A Semi-supervised Nighttime Dehazing Baseline with Spatial-Frequency Aware and Realistic Brightness Constraint. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 631–640.

[6] Yuning Cui, Wenqi Ren, Xiaochun Cao, and Alois Knoll. 2024. Revitalizing convolutional network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 12 (2024), 9423–9438.

[7] Xin Dong, Barbara De Salvo, Meng Li, Chiao Liu, Zhongnan Qu, Hsiang-Tsung Kung, and Ziyun Li. 2022. Splitnets: Designing neural architectures for efficient distributed computing on head-mounted systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12559–12569.

[8] Chengyu Fang, Chunming He, Fengyang Xiao, Yulun Zhang, Longxiang Tang, Yuelin Zhang, Kai Li, and Xiu Li. 2024. Real-world image dehazing with coherence-based pseudo labeling and cooperative unfolding network. *Neural Information Processing Systems* 37 (2024), 97859–97883.

[9] Jiayi Fu, Siyu Liu, Zikun Liu, Chun-Le Guo, Hyunhee Park, Ruiqi Wu, Guoqing Wang, and Chongyi Li. 2025. Iterative Predictor-Critic Code Decoding for Real-World Image Dehazing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 12700–12709.

[10] Yulu Gan, Mingjie Pan, Rongyu Zhang, Zijian Ling, Lingran Zhao, Jiaming Liu, and Shanghang Zhang. 2023. Cloud-device collaborative adaptation to continual changing environments in the real-world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12157–12166.

[11] Chen Gong, Rui Xing, Zhenzhe Zheng, and Fan Wu. 2025. A two-stage data selection framework for data-efficient model training on edge devices. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 673–684.

[12] Jie Gui, Xiaofeng Cong, Yuan Cao, Wenqi Ren, Jun Zhang, Jing Zhang, Jiuxin Cao, and Dacheng Tao. 2023. A comprehensive survey and taxonomy on single image dehazing based on deep learning. *Comput. Surveys* 55, 13s (2023), 1–37.

[13] Siyan Guo, Cong Zhao, Shusen Yang, Yingying Liang, Yimeng Wang, and Qing Han. 2025. Edge-Cloud Collaborative Real-Time Video Object Detection for Industrial Surveillance Systems. *IEEE Intelligent Systems* 40 (2025), 55–63.

[14] Shaohui Jin, Ziqin Xu, Mingliang Xu, and Hao Liu. 2024. Time-gated imaging through dense fog via physics-driven swin transformer. *Optics Express* 32, 11 (2024), 18812–18830.

[15] Pirunthan Keerthinathan, Narmilan Amarasingam, Grant Hamilton, and Felipe Gonzalez. 2023. Exploring unmanned aerial systems operations in wildfire management: Data types, processing algorithms and navigation. *International Journal of Remote Sensing* 44, 18 (2023), 5628–5685.

[16] Connor Lee, Matthew Anderson, Nikhil Ranganathan, Xingxing Zuo, Kevin Do, Georgia Gkioxari, and Soon-Jo Chung. 2024. Caltech Aerial RGB-Thermal Dataset in the Wild. In *European Conference on Computer Vision*. 236–256.

[17] Jing Li, Lu Bai, Bin Yang, Chang Li, and Lingfei Ma. 2025. Graph representation learning for infrared and visible image fusion. *IEEE Transactions on Automation Science and Engineering* 22 (2025), 13801–13813.

[18] Junjie Li, Kaichen Chi, Yue Chang, and Qi Wang. 2025. DFG-DDM: Deep frequency-guided denoising diffusion model for remote sensing image dehazing. *IEEE Transactions on Geoscience and Remote Sensing* 63 (2025), 1–12.

[19] Jiawei Li, Hongwei Yu, Jiansheng Chen, Xinlong Ding, Jinlong Wang, Jinyuan Liu, Bochao Zou, and Huimin Ma. 2025. A^2RNet: Adversarial Attack Resilient Network for Robust Infrared and Visible Image Fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 4770–4778.

[20] Ke Li, Di Wang, Zhangyuan Hu, Shaofeng Li, Weiping Ni, Lin Zhao, and Quan Wang. 2025. Fd2-net: Frequency-driven feature decomposition network for infrared-visible object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 4797–4805.

[21] Yanyu Li, Geng Yuan, Yang Wen, Ju Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang, and Jian Ren. 2022. Efficientformer: Vision transformers at mobilenet speed. *Advances in Neural Information Processing Systems* 35 (2022), 12934–12949.

[22] Yudong Liang, Bin Wang, Wangmeng Zuo, Jiaying Liu, and Wenqi Ren. 2022. Self-supervised Learning and Adaptation for Single Image Dehazing.. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 1137–1143.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*. Springer, 740–755.

[24] Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. 2023. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8115–8124.

[25] Jinyuan Liu, Guanyao Wu, Zhu Liu, Di Wang, Zhiying Jiang, Long Ma, Wei Zhong, and Xin Fan. 2024. Infrared and visible image fusion: From data compatibility to task adaption. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47, 4 (2024), 2349–2369.

[26] Jinyuan Liu, Bowei Zhang, Qingyun Mei, Xingyuan Li, Yang Zou, Zhiying Jiang, Long Ma, Risheng Liu, and Xin Fan. 2025. DCEvo: Discriminative Cross-Dimensional Evolutionary Learning for Infrared and Visible Image Fusion. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*. 2226–2235.

[27] Xinyu Liu, Houwen Peng, Ningxin Zheng, Yuqing Yang, Han Hu, and Yixuan Yuan. 2023. Efficientvit: Memory efficient vision transformer with cascaded group attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14420–14430.

[28] Zhu Liu, Jinyuan Liu, Benzhuang Zhang, Long Ma, Xin Fan, and Risheng Liu. 2023. PAIF: Perception-aware infrared-visible image fusion for attack-tolerant semantic segmentation. In *Proceedings of the ACM International Conference on Multimedia*. 3706–3714.

[29] Hu Lu, Xuezhang Zou, and Pingping Zhang. 2023. Learning progressive modality-shared transformers for effective visible-infrared person re-identification. In *Proceedings of the AAAI conference on Artificial Intelligence*, Vol. 37. 1835–1843.

[30] Xianting Lu, Yunong Wang, Yu Fu, Qi Sun, Xuhua Ma, Xudong Zheng, and Cheng Zhuo. 2024. MISP: A Multimodal-based Intelligent Server Failure Prediction Model for Cloud Computing Systems. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5509–5520.

[31] Long Ma, Yuxin Feng, Yan Zhang, Jinyuan Liu, Weimin Wang, Guang-Yong Chen, Chengpei Xu, and Zhuo Su. 2025. Coa: Towards real image dehazing via compression-and-adaptation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 11197–11206.

[32] Yoshitomo Matsubara, Matteo Mendula, and Marco Levorato. 2025. A Multi-Task Supervised Compression Model for Split Computing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 4913–4922.

[33] Purva Makarand Mhasakar, Kevin Doshi, Ning Wang, Shen Shyang Ho, and Haibin Ling. 2023. Distributed Tracking and Verifying: A Real-Time and High-Accuracy Visual Tracking Edge Computing Framework for Internet of Things. In *Proceedings of the ACM/IEEE Symposium on Edge Computing*. 360–364.

[34] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. 2012. Making a completely blind image quality analyzer. *IEEE Signal Processing Letters* 20, 3 (2012), 209–212.

[35] Liuxiang Qiu, Si Chen, Yan Yan, Jing-Hao Xue, Da-Han Wang, and Shunzhi Zhu. 2024. High-order structure based middle-feature learning for visible-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 4596–4604.

[36] Yuwei Qiu, Kaihao Zhang, Chenxi Wang, Wenhan Luo, Hongdong Li, and Zhi Jin. 2023. MB-TaylorFormer: Multi-branch efficient transformer expanded by taylor formula for image dehazing. In *Proceedings of the International Conference on Computer Vision*. 12802–12813.

[37] Xiaoyi Qu, David Aponte, Colby Banbury, Daniel P Robinson, Tianyu Ding, Kazuhito Koishida, Ilya Zharkov, and Tianyi Chen. 2025. Automatic joint structured pruning and quantization for efficient neural network training and compression. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 15234–15244.

[38] Kathakoli Sengupta, Zhongkai Shangguan, Sandesh Bharadwaj, Sanjay Arora, Eshed Ohn-Bar, and Renato Mancuso. 2024. Unified Local-Cloud Decision-Making via Reinforcement Learning. In *European Conference on Computer Vision*. Springer, 185–203.

[39] Junseong Shin, Seungwoo Chung, Yunjeong Yang, and Tae Hyun Kim. 2025. HazeFlow: Revisit Haze Physical Model as ODE and Non-Homogeneous Haze Generation for Real-World Dehazing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6263–6272.

[40] Mehrdad Khani Shirkoohi, Pouya Hamadanian, Arash Nasr-Esfahany, and Mohammad Alizadeh. 2021. Real-Time Video Inference on Edge Devices via Adaptive Model Streaming.. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Vol. 21. 4552–4562.

[41] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. 2023. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing* 32 (2023), 1927–1941.

[42] Huixin Sun, Runqi Wang, Yanjing Li, Linlin Yang, Shaohui Lin, Xianbin Cao, and Baochang Zhang. 2025. SET: Spectral Enhancement for Tiny Object Detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 4713–4723.

[43] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. 2022. Detfusion: A detection-driven infrared and visible image fusion network. In *Proceedings of the ACM International Conference on Multimedia*. 4003–4011.

[44] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. 2022. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 10 (2022), 6700–6713.

[45] Linfeng Tang, Xinyu Xiang, Hao Zhang, Meiqi Gong, and Jiayi Ma. 2023. DIVFusion: Darkness-free infrared and visible image fusion. *Information Fusion* 91 (2023), 477–493.

[46] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, and Chia-Wen Lin. 2025. PHATNet: A Physics-guided Haze Transfer Network for Domain-adaptive Real-world Image Dehazing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5591–5600.

[47] Kashif Usmani, Timothy O'Connor, Pranav Wani, and Bahram Javidi. 2022. 3D object detection through fog and occlusion: passive integral imaging vs active (LiDAR) sensing. *Optics Express* 31, 1 (2022), 479–491.

[48] Rejin Varghese and M Sambath. 2024. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *International Conference on Advances in Data Engineering and Intelligent Computing Systems*. IEEE, 1–6.

[49] Narasimhan Venkatanath, D Praneeth, Maruthi Chandrasekhar Bh, Sumohana S Channappayya, and Swarup S Medasani. 2015. Blind image quality evaluation using perception based features. In *National Conference on Communications*. 1–6.

[50] Cong Wang, Jinshan Pan, Wanyu Lin, Jiangxin Dong, Wei Wang, and Xiao-Ming Wu. 2024. SelfPromer: Self-prompt dehazing transformers with depth-consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 5327–5335.

[51] Cong Wang, Jinshan Pan, Liyan Wang, and Wei Wang. 2025. Intra and Inter Parser-Prompted Transformers for Effective Image Restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 7609–7618.

[52] Chuanming Wang, Yuxin Yang, Mengshi Qi, Huanhuan Zhang, and Huadong Ma. 2025. Towards Efficient Object Re-Identification with a Novel Cloud-Edge Collaborative Framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 7600–7608.

[53] Guanqun Wang, Jiaming Liu, Chenxuan Li, Yuan Zhang, Junpeng Ma, Xinyu Wei, Kevin Zhang, Maurice Chong, Renrui Zhang, Yijiang Liu, et al. 2024. Cloud-device collaborative learning for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12646–12655.

[54] Maolin Wang, Jun Chu, Sicong Xie, Xiaoling Zang, Yao Zhao, Wenliang Zhong, and Xiangyu Zhao. 2025. Put Teacher in Student's Shoes: Cross-Distillation for Ultra-compact Model Compression Framework. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4975–4985.

[55] Ruiyi Wang, Yushuo Zheng, Zicheng Zhang, Chunyi Li, Shuaicheng Liu, Guangtao Zhai, and Xiaohong Liu. 2025. Learning Hazing to Dehazing: Towards Realistic Haze Generation for Real-World Image Dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23091–23100.

[56] Xiao Wang, Lekai Liu, Bin Yang, Mang Ye, Zheng Wang, and Xin Xu. 2025. TokenMatcher: Diverse Tokens Matching for Unsupervised Visible-Infrared Person Re-Identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 7934–7942.

[57] Xinran Wang, Guang Yang, Tian Ye, and Yun Liu. 2025. Dehaze-RetinexGAN: Real-World Image Dehazing via Retinex-based Generative Adversarial Network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39(8). 7997–8005.

[58] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612.

[59] Jing Wu, Peng Wei, and Feng Huang. 2024. Color-preserving visible and near-infrared image fusion for removing fog. *Infrared Physics & Technology* 138 (2024), 105252.

[60] Likang Wu, Zhi Li, Hongke Zhao, Zhefeng Wang, Qi Liu, Baoxing Huai, Nicholas Jing Yuan, and Enhong Chen. 2023. Recognizing unseen objects via multimodal intensive knowledge graph propagation. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2618–2628.

[61] Xiaoyu Xia, Sheik Mohammad Mostakim Fattah, and Muhammad Ali Babar. 2023. A survey on UAV-enabled edge computing: Resource management perspective. *Comput. Surveys* 56, 3 (2023), 1–36.

[62] Tianyang Xu, Yifan Pan, Zhenhua Feng, Xuefeng Zhu, Chunyang Cheng, Xiao-Jun Wu, and Josef Kittler. 2024. Learning Feature Restoration Transformer for Robust Dehazing Visual Object Tracking. *International Journal of Computer Vision* (2024), 1–18.

[63] Jianzhong Yang, Xiaoqing Ye, Bin Wu, Yanlei Gu, Ziyu Wang, Deguo Xia, and Jizhou Huang. 2022. DuARE: Automatic road extraction with aerial images and trajectory data at Baidu maps. In *Proceedings of the ACM SIGKDD conference on knowledge discovery and data mining*. 4321–4331.

[64] Run Yang, Hui He, Yixiao Xu, Bangzhou Xin, Yulong Wang, Yue Qu, and Weizhe Zhang. 2023. Efficient intrusion detection toward IoT networks using cloud–edge collaboration. *Computer Networks* 228 (2023), 109724.

[65] Tian Ye, Sixiang Chen, Haoyu Chen, Wenhao Chai, Jingjing Ren, Zhaohu Xing, Wenxue Li, and Lei Zhu. 2025. PromptHaze: Prompting Real-world Dehazing via Depth Anything Model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 9454–9462.

[66] Meng Yu, Te Cui, Haoyang Lu, and Yufeng Yue. 2024. VIFNet: An end-to-end visible-infrared fusion network for image dehazing. *Neurocomputing* (2024), 128105.

[67] Yazhou Yuan, Shicong Gao, Ziteng Zhang, Wenye Wang, Zhezhuang Xu, and Zhixin Liu. 2024. Edge-cloud collaborative UAV object detection: Edge-embedded lightweight algorithm design and task offloading using fuzzy neural network. *IEEE Transactions on Cloud Computing* 12, 1 (2024), 306–318.

[68] Zhongzheng Yuan, Samyak Rawlekar, Siddharth Garg, Elza Erkip, and Yao Wang. 2024. Split computing with scalable feature compression for visual analytics on the edge. *IEEE Transactions on Multimedia* 26 (2024), 10121–10133.

[69] Jianshan Zhang, Haibo Luo, Xing Chen, Hong Shen, and Longkun Guo. 2024. Minimizing Response Delay in UAV-Assisted Mobile Edge Computing by Joint UAV Deployment and Computation Offloading. *IEEE Transactions on Cloud Computing* 12, 4 (2024), 1372–1386.

[70] Kaili Zhang, Anzhi Wang, Yawei Xiong, and Yun Liu. 2024. Survey of Transformer-Based Single Image Dehazing Methods. *Journal of Frontiers of Computer Science & Technology* 18, 5 (2024).

[71] Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. 2022. Visible-thermal UAV tracking: A large-scale benchmark and new baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8886–8895.

[72] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 586–595.

[73] Xingchen Zhang and Yiannis Demiris. 2023. Visible and infrared image fusion using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 8 (2023), 10535–10554.

[74] Yafei Zhang, Yu Liu, et al. 2026. Infrared-Assisted Single-Stage Framework for Joint Restoration and Fusion of Visible and Infrared Images under Hazy Conditions. *Pattern Recognition* (2026), 113074.

[75] Junwei Zhao, Qianchun Luo, Shiliang Zhang, Shen Gao, and Jie Wu. 2025. HDCFN: Haze Distribution-aware Cross-modal Fusion Network for Infrared-guided Dense Haze Removal in UAVs. In *Proceedings of the ACM International Conference on Multimedia*. 10867–10875.

[76] Wenda Zhao, Shigeng Xie, Fan Zhao, You He, and Huchuan Lu. 2023. Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13955–13965.

[77] Muhammad Aidiel Zulkifley, Mehran Behjati, Rosdiadee Nordin, and Mohamad Shanudin Zakaria. 2021. Mobile Network Performance and Technical Feasibility of LTE-Powered Unmanned Aerial Vehicle. *Sensors* 21, 8 (2021), 2848.