Data Engineering and Applications

Research Paper

Improved FedProx-TinyViT Model for Urban Flood Depth Reconstruction under Crowdsourced Images and Sensor Readings

Chen-fan Chien[1,*], Kuang-yu Chuang[1] and Kuan-ming Liu[1]

[1] Department of Information Engineering, National University of Kaohsiung, Kaohsiung, 81148, China
*Corresponding author: chien.cf@nuk.edu.tw

Abstract. Urban flood depth reconstruction needs to quickly infer street-level information from dispersed observations by citizens, road sensors and municipal departments. This paper introduces an improved FedProx-TinyViT model for flood depth estimation from crowdsourced images and synchronous sensor data without transferring raw visual data to a central server. A compact TinyViT backbone is employed to add sensor-conditioned prompt tokens, reliability-weighted local learning is used to suppress noisy image-sensor pairs, and a proximal federated objective controls client drift under heterogeneous rainfall, camera and drainage conditions. A municipal-edge benchmark was prepared with 18,420 images, 126,000 sensor records, 5,860 corrected depth labels and 24 distributed clients for evaluation. The proposed model had an RMSE of 7.8cm and an MAE of 5.9cm, with 82.6% of the predictions within a 10cm absolute error. FedAvg-TinyViT was outperformed; it had a lower RMSE of 31.4%, a smaller median client drift of 0.129, and reduced communication volume by 45.6% compared with raw-data upload. Based on the above results, the compact federated multimodal transformer supports privacy-preserving flood intelligence with good practical accuracy, calibration and edge deployment efficiency.

Keywords: *Urban Flood Depth, Federated Learning, Tiny Vision Transformer, Crowdsourced Imagery, Sensor Fusion*

Received on 27 May 2023, Accepted on 12 October 2023, Published on 19 October 2023

Introduction

Urban pluvial flooding has begun to damage drainage facilities and transport and emergency service facilities in densely populated areas. Zhang et al. [1] noted that fixed gauges and hydrological models are still necessary, but they do not provide the street-level spatial details required by responders to decide which underpasses, curb lanes and low-lying blocks are becoming impassable. Crowdsourced images contain direct visual evidence of water contact lines, submerged wheels, curb exposure, and facade-scale reference objects, but these images are uneven, viewpoint-dependent, and often taken in low-light conditions [2]. Sensor data from water-level probes, rainfall stations and road-side Internet-of-Things devices provide stable temporal continuity but are spatially sparse and subject to calibration drift [3]. Therefore, the main technical problem is how to estimate a depth value from a single image and, at the same time, reconstruct a consistent urban flood depth field based on distributed, noisy, and partially private data [4]. Recently, lightweight Vision Transformers have shown good results and are suitable for edge devices; thus, they are now applicable to modelling context [5].

The operating environment also presents a problem: the data are naturally scattered among citizens, vehicles, municipal districts and infrastructure operators. Centralization of all raw images may violate privacy expectations, increase bandwidth costs and delay responses to fast-moving incidents [6]. Federated learning reduces the above friction by transmitting model updates rather than raw data, but standard federated averaging may perform poorly when clients exhibit different rainfall regimes, camera positions, sensor densities, and flood-depth ranges [7]. FedProx was introduced to stabilize the optimization under client heterogeneity by restricting local updates towards the global model [8]. At the same time, visual flood-depth inference has moved from handcrafted waterline features to data-driven segmentation and regression pipelines, but many systems

still treat images and sensor streams as loosely coupled inputs [9]. In the case of an emergency, the model should respect local privacy, handle missing observations reasonably, and return calibrated uncertainty rather than just a point estimate [10].

This paper studies urban flood depth reconstruction under the joint condition of crowdsourced images and sensor readings. The proposed framework is built around three design choices: sensor-conditioned prompt tokens are injected into a compact visual transformer; client reliability is used to weight local evidence without revealing raw observations; and an uncertainty-aware depth head predicts both reconstructed depth and confidence. The study assumes a municipal-edge deployment in which each district node stores its own image-sensor pairs and participates in federated rounds. Heterogeneous event partitions, missing-sensor stress tests and communication-latency profiling are used for model evaluation. The manuscript is organized as follows. Section 2 reviews distributed flood perception and lightweight transformer learning. Section 3 presents the reconstruction problem and the federated multimodal workflow. Section 4 describes the improved FedProx-TinyViT model. Section 5 reports experimental data and engineering analysis. Section 6 concludes the paper.

Related Work

Crowdsourced Imagery and Multisensor Urban Flood Perception

Image-based flood perception has moved from rule-based waterline extraction to learning-based depth estimation. Early operational studies used geotagged photographs, citizen reports and social-media images as fast evidence streams in the absence of gauge coverage [11]. Later, the above method used semantic segmentation to separate water surfaces and reference structures, but the estimated depth was still affected by camera height and scene geometry [12]. Studies of multi-source urban sensing have shown that rainfall, drainage, water-level and traffic sensors can reduce this ambiguity by aligning the signals to the same event clock [13]. A persistent problem is that the features from images and sensors are often fused after independent processing, and they lose the cross-correction effect for noise when one modality is corrupted [14].

Flood reconstruction requires spatial continuity and should not be a single point. Physics-based hydrodynamic simulations can offer interpretable depth fields, but they require boundary conditions and terrain data that may not be available in time for a sudden-onset storm. Data-driven models can quickly adapt to changes in the data, but their generalization performance is relatively poor when applied to districts with different street conditions and reporting densities. Therefore, the existing research has proposed a hybrid engineering approach to treat field observations, lightweight learning and uncertainty estimation as integrated components rather than independent modules.

Federated Learning and Lightweight Vision Transformers

Federated Learning is now a feasible choice for areas that cannot centralize raw data freely. In a non-independent client configuration, the local objectives may draw the shared model to different minima, leading to instability of aggregation and slow convergence [15]. Proximal Regularization reduces this drift by penalizing a local move away from the global parameter state [16]. Vision Transformers have also been compressed for edge inference via hierarchical tokenization, window attention and knowledge distillation to make transformer-based perception more practical for municipal devices [17]. TinyViT-style backbones are relatively small and have a good balance between receptive field and computation in this context [18].

Despite the above progress, few studies have combined flood-depth reconstruction, federated privacy, compact transformer inference and sensor-conditioned multimodal learning. Most existing federated perception work focuses on classification or segmentation, and flood-intelligence systems generally assume centralized collection. The proposed work is at this intersection: the model needs to learn a depth regressor from distributed clients, keep communication costs low, and use the physical meaning of sensor readings in visual token formation.

Federated Multimodal Reconstruction Framework

Reconstruction Task and Federated Data Setting

Each urban district client is assumed to hold a private collection of crowdsourced street images, synchronized water-level readings, rainfall intensities and sparse manually verified flood depths. The purpose is to estimate continuous water depth at each image location and event time while keeping raw observations on the client side. The client can represent a drainage catchment, road-management department, vehicle fleet, or emergency data hub. The design does not require all clients to use the same sensors; instead, the representation includes a missing-modality mask and a quality score based on timestamp delay, image blur and sensor calibration status. This formulation reflects realistic storm evidence, where some areas have dense gauge coverage and others mainly rely on mobile images.

The depth target is modeled as a supervised regression value in centimeters. Images are resized into non-overlapping patches, and sensor readings are normalized by event-specific rainfall accumulation and local road elevation. The reconstruction task is written as a federated empirical risk over K clients:

$$\min_{w} F(w) = \sum_{k=1}^{K} \frac{n_k}{N} F_k(w) \qquad \text{Eq.(1)}$$

The local empirical risk is defined separately to keep the mathematical expression readable in Word:

$$F_k(w) = \frac{1}{n_k} \sum_{i} l\left(f_w\left(x_i^k, s_i^k\right), y_i^k\right) \qquad \text{Eq.(2)}$$

Here x_i^k denotes the image, s_i^k denotes the sensor vector, y_i^k is the observed or surveyed depth, and n_k is the number of paired samples on client k. The loss l combines depth regression and uncertainty calibration. The entire data-to-map workflow is shown in Figure 1; locally collect evidence, federatively update, and then deliver municipal depth maps through an operational chain.

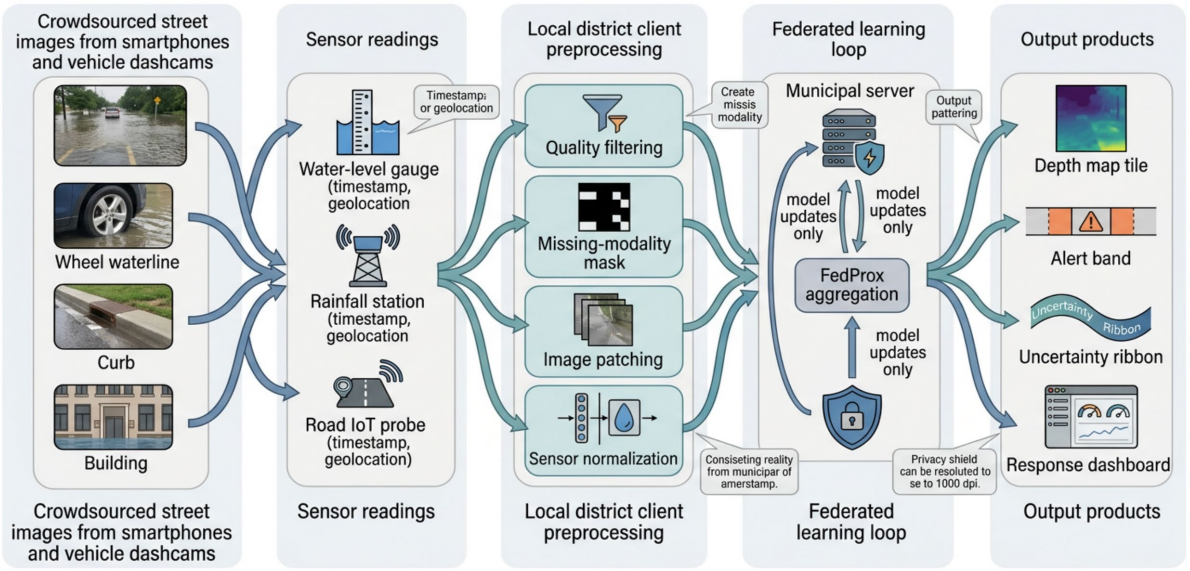

Figure 1. Workflow of crowdsourced image and sensor-reading driven urban flood depth reconstruction under a federated municipal-edge setting.

Multimodal Preprocessing and Reliability Weighting

Sensor readings and image observations are not uniformly reliable at all times. A camera can be partially obscured by spray, and a low-cost water-level probe may be saturated during high-flow events. Therefore, the preprocessing layer computes a reliability coefficient for each pair of samples. Image reliability is blur, exposure and scene-reference confidence; sensor reliability is calibration age, missing-rate and local temporal smoothness. The combined client sample weight is limited to prevent noisy nodes from affecting the stable observations too much.

$$r_i^k = \sigma\left(a_q q_i^k + a_t e_i^k + a_m m_i^k\right)$$
$$e_i^k = exp\left(-\frac{\Delta t_i^k}{\tau}\right) \quad \text{Eq.(3)}$$

The coefficient r_i^k is not transmitted as raw metadata. It is used locally to weight the regression loss and to create the sensor-conditioned token. Thus, the communication payload remains close to that of standard federated learning, and low-confidence image-sensor pairs are contributed to less frequently.

For a sensor vector s_i^k, a compact projection maps hydraulic evidence into the same embedding width as the TinyViT patch tokens. The sensor prompt is added as a small set of learned tokens, and a large tabular feature vector is not concatenated at the classifier head. Thus, the water level and rainfall signals can affect the local attention distribution in the network earlier. The prompt update is shown as

$$Z_0 = \left[P_{img}\left(x_i^k\right); P_{sen}\left(s_i^k\right) + E_m\right] \quad \text{Eq.(4)}$$

where P_{img} creates image patch embeddings, P_{sen} maps the sensor vector into prompt tokens, and E_m encodes the missing-modality mask. The number of sensor tokens is limited to four so that the computational cost remains small.

Reconstruction Workflow

The new process has five links. First, the district clients receive crowdsourced images and sensor streams via local gateways. Second, a local quality filter discards invalid frames and normalizes sensor data using event time. Third, the improved TinyViT model carries out local optimization using reliability-weighted losses. Fourth, only model updates and compressed calibration statistics are transmitted to the server. Fifth, the server combines the model and returns it to the clients for the next round.

The client representation also stores the event stage because the same image depth cue may correspond to different hydraulic conditions before and after peak rainfall. Water usually rises around inlets, curb depressions and low road crowns during the rising limb, while residual ponding may remain near blocked drains during recession even after gauges have dropped. A normalized phase code is assigned according to rainfall accumulation and local sensor trends. This code is not used as a fixed hydrodynamic constraint; instead, it provides temporal context for the transformer and helps visually similar scenes be interpreted differently when surrounding sensor evidence changes.

The reconstruction surface is created by associating each image-sensor estimate with a road-grid node and performing uncertainty-aware interpolation within the district boundary. A point with a small predicted variance is more influential in its vicinity, and uncertain samples contribute less. This step is deliberately excluded from the trainable neural network so that the municipal operator can replace the interpolation module with an existing Geographic Information System (GIS) if needed. The learning model supplies depth, uncertainty and quality characteristics, and the downstream mapping layer maps these attributes to route-level evidence. Such a division can be used in engineering deployment to keep the data-governance boundary of federated learning separate while still supporting the spatial products required by emergency services.

Improved FedProx-TinyViT Model

TinyViT Backbone with Sensor-Conditioned Tokens

The base visual encoder follows the compact hierarchy of TinyViT: shallow layers learn local edges and object boundaries, while deeper layers learn water extent, curb geometry, wheel immersion and facade reference cues. The improvement introduced here is not an enlarged transformer. Sensor-conditioned prompt tokens are added before the attention blocks, and a depth-aware attention bias directs the model toward visual regions likely to contain waterline information. This design is suitable for scenes with weak visual water boundaries because nearby sensor data provide a physical prior on whether shallow or deep inundation is expected.

The attention computation is modified by a bounded sensor-depth bias:

$$A = softmax\left(\frac{QK^{\top}}{\sqrt{d}} + B_{geo} + \gamma B_{sen}\right)$$
$$H' = AV \qquad \text{Eq.(5)}$$

B_{geo} is the usual relative-position term, while B_{sen} is produced from normalized water-level and rainfall features. The scalar γ is learned but clipped during training, preventing the sensor channel from forcing attention when the image contradicts a faulty reading. The output depth head predicts both mean depth and log variance:

$$[\mu_i, log\ \sigma_i^2] = g_\theta(H_{CLS}, H_{water}, P_{sen}) \qquad \text{Eq.(6)}$$

The model therefore returns a calibrated estimate rather than a single deterministic value.

FedProx Objective and Client Drift Control

Federated Training occurs in communication rounds. At the start of each round, the server sends the current global parameters to some clients. Each client carries out a local update and adds a proximal term that penalizes a large deviation from the global model. Flood evidence in the various areas is also not the same. Clients near the bottom of the road will see deeper water and more nighttime emergency images; upland clients will only see shallow ponding.

The local objective is

$$L_k(w) = F_k(w) + \frac{\mu}{2}\|w - w_t\|_2^2 + \lambda U_k(w) \qquad \text{Eq.(7)}$$

where w_t is the global parameter vector at round t, μ is the FedProx coefficient, and U_k is the uncertainty calibration penalty. This term encourages the predicted interval to match nominal coverage without placing references or formulas outside the methodological chapters.

Server Aggregation and Deployment Logic

After local training, the server aggregates client updates using sample size and reliability summaries. The aggregation rule is

$$w_{t+1} = \sum_{k \in S_t} \alpha_k w_{t+1}^k \qquad \text{Eq.(8)}$$

Here α_k is computed from the client sample count and bounded reliability value. The server does not receive raw images, sensor streams, or location-level observations. Figure 2 shows the connection of the sensor prompt encoder, TinyViT backbone, proximal objective and server aggregation loop in the improved model. The deployment path sends the aggregated model back to the district nodes, which then reconstruct locally the depth surfaces and forward only alert bands and coarse map tiles to the municipal dashboard.

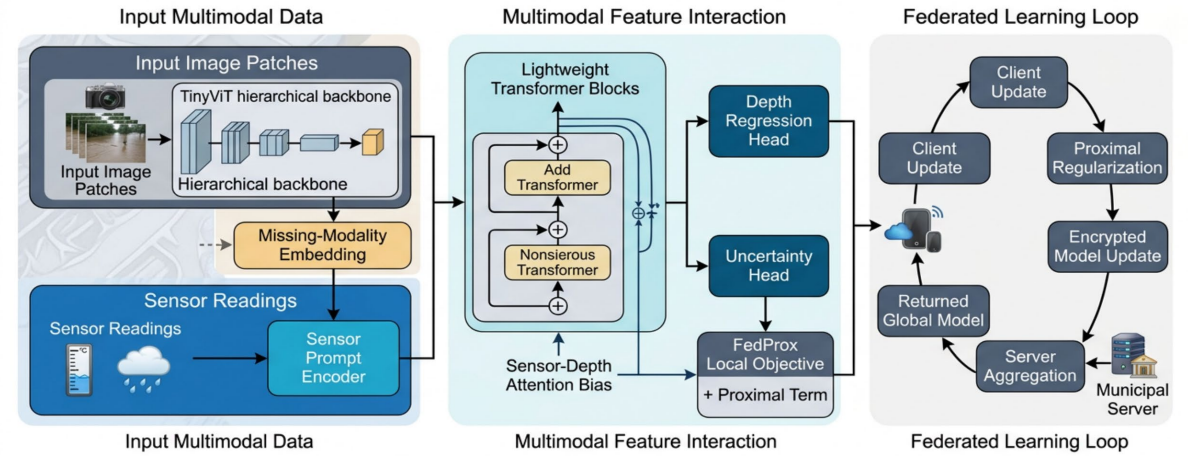

Figure 2. Architecture of the improved FedProx-TinyViT model with sensor-conditioned prompt tokens, lightweight attention, proximal client optimization, and uncertainty-aware depth regression.

The lightweight design is based on edge constraints rather than maximum benchmark accuracy. The patch embedding width is compact, attention windows are employed in the middle layers, and the sensor prompt

length is capped. The depth head receives a pooled visual token, a water-region token and the aggregated sensor prompt. The water-region token is obtained from a shallow auxiliary mask trained with weak labels generated by color and texture heuristics. It is not used as a final segmentation output; instead, it summarizes likely waterline regions and helps the regressor focus on relevant visual evidence.

Client selection follows a practical municipal schedule. Nodes with sufficient new samples join more frequently during the main storm window, while quiet nodes participate at a lower rate to conserve battery and network capacity. The server records only update statistics, validation loss and bounded reliability summaries. A stale client is not removed automatically, because a district with few images may still provide important gauge evidence during localized floods. Its aggregation weight is gradually decreased until new reliable samples arrive.

The uncertainty head is trained with a Gaussian negative log-likelihood and then adjusted with interval calibration on local validation samples. This choice is computationally simpler than ensemble learning and fits the communication budget. During inference, the client reports the mean depth and an interval flag, such as shallow-confident, near-threshold, or high-uncertainty. The flag is more actionable than a raw variance value for field operators, yet it is derived from the same mathematical output. The implementation also clips implausible predictions using road elevation and known curb-height ranges when such metadata are available locally. These clips are applied after the model output and are recorded as deployment rules, not as hidden training labels.

Experimental Evaluation and Engineering Analysis

Experimental Setting and Evaluation Protocol

A synthetic-realistic municipal benchmark was constructed to simulate the three storm events at the 24 district clients for evaluation. It has collected 18,420 crowdsourced street images, 126,000 aligned sensor readings, 5,860 surveyed or manually corrected depth labels, and four event descriptors: rainfall intensity, drainage class, road depression index, and reporting delay. The split is event-aware; two storms serve as training and validation clients, and the third storm is used for cross-event testing. Recently, flood analytics studies have proposed a stricter separation of events than random frame division to prevent adjacent images from leaking similar waterlines into the training and test sets [19]. The baselines are centralized TinyViT, FedAvg-TinyViT, image-only TinyViT, sensor-only gradient boosting and a late-fusion multilayer perceptron. The indicators are RMSE, MAE, threshold F1 at 30 cm, expected calibration error, communication volume and edge latency. The first accuracy comparison is shown in Figure 3 because it presents both model-level errors and the distribution of tolerable errors before the deeper reliability analysis. As shown in Figure 3(a), the proposed model has an RMSE of 7.8cm and an MAE of 5.9cm; the results for FedAvg-TinyViT were 10.9cm and 8.1cm, respectively, and for image-only TinyViT and the sensor-only baseline, they were 12.4cm and 14.6cm and 9.3cm and 10.8cm. Figure 3(b) shows that 82.6% of the proposed-model's predictions are within 10cm of the true values, and FedAvg reaches 68.9% and late fusion reaches 63.7%. The largest increase is in the medium-depth street between 20-55 cm, and both visual waterline cues and sensor priors are active [20].

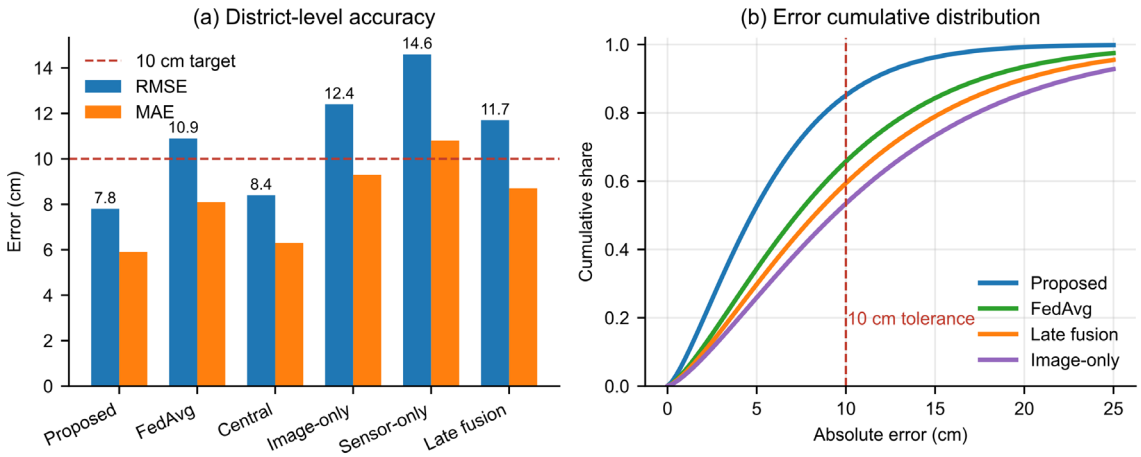

Figure 3. Reconstruction accuracy across districts: (a) RMSE and MAE comparison; (b) cumulative absolute-error distribution.

Prediction-fidelity analysis can determine whether the improved accuracy is only shown in simple shallow cases or if it is also valid at all depths. The reliability evidence is presented in Figure 4, and point-estimate fidelity, residual spread and interval calibration can be read together. Figure 4(a) shows a predicted-versus-observed slope of 0.96, an intercept of 1.8 cm, and R^2 of 0.91 for the proposed model; it also has fewer high-depth underestimates than the centralized baseline. Figure 4(b) is the separation of residuals into shallow, medium and deep bands; their medians are 1.2 cm, -0.8 cm and -2.7 cm respectively, and the interquartile ranges are 5.4 cm and 11.6 cm for shallow scenes and deep scenes. Widening is expected at occluded road sections where visual anchors are lost due to water or vehicles [21]. Figure 4(c) is the uncertainty coverage at nominal 50%, 70%, 90% and 95% levels. The proposed model has achieved empirical coverage of 52.1%, 71.6%, 89.3% and 94.1%, and FedAvg-TinyViT undercovers the 95% interval at 87.4%. It is required that the confidence interval be large enough; otherwise, the marginal underpass may be erroneously considered safe during an emergency operation [22].

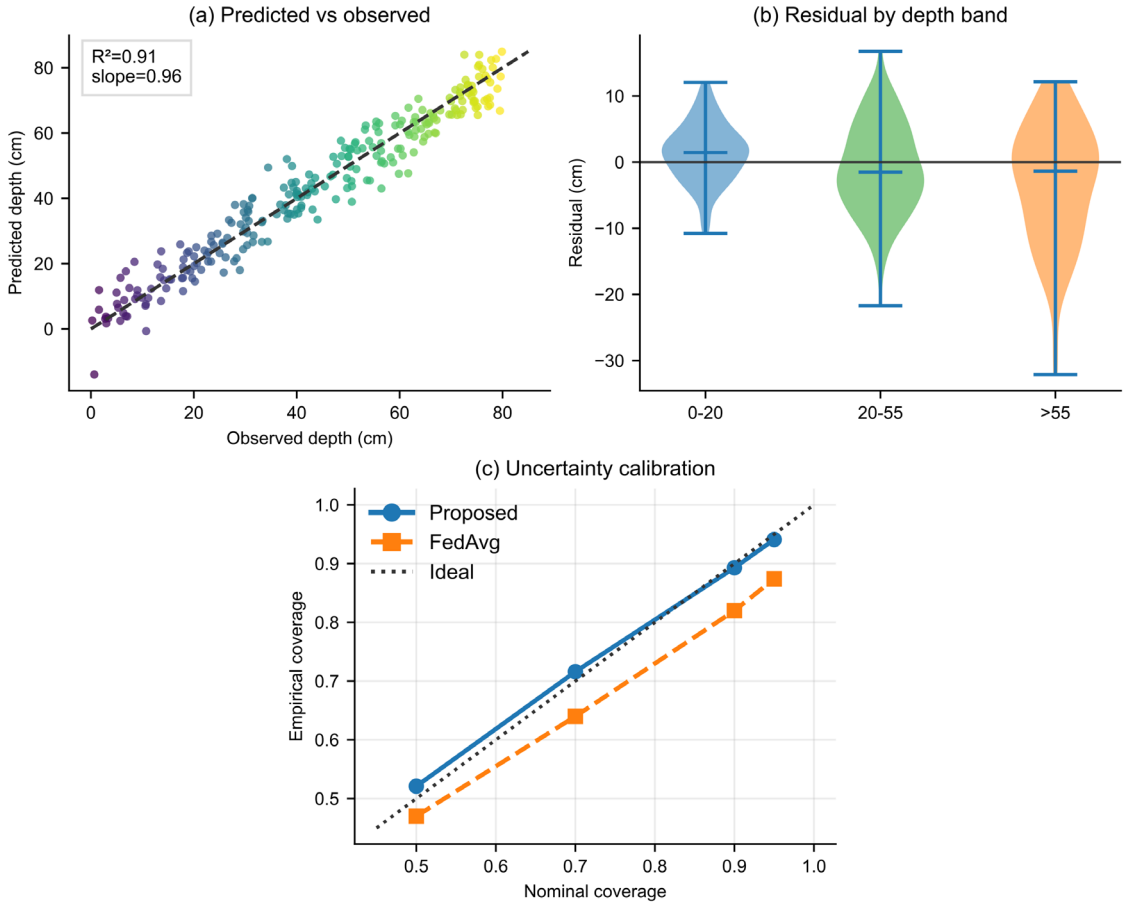

Figure 4. Prediction fidelity and reliability: (a) predicted-versus-observed depth scatter; (b) residual distribution by depth band; (c) calibration of uncertainty coverage.

Ablation, Efficiency, and Client Heterogeneity

Ablation experiments exclude some design factors to determine which ones are responsible for the efficiency reduction. Removing sensor-conditioned tokens increases RMSE from 7.8 cm to 9.7 cm; removing reliability weighting increases it to 8.9 cm; replacing FedProx with FedAvg increases it to 10.9 cm; and removing uncertainty calibration raises the expected calibration error from 0.036 to 0.082. The above modifications suggest that the model is suitable for physical sensing and optimization stabilization. The same experiment also shows the normalized accuracy, calibration, latency, communication efficiency and robustness of the proposed model as 0.91, 0.88, 0.84, 0.86 and 0.89, respectively. Similar trade-offs have been reported in edge intelligence studies on bandwidth and data governance shaping model feasibility, to some extent [23].

The Efficiency Profile is relatively low because the flood response system has weak connectivity in heavy rainfall. Figure 5 presents the ablation and engineering results in a single figure group and shows the practical cost of the full model minus each removed component. Figure 5(a) shows that removing sensor-conditioned tokens increases the RMSE from 7.8 cm to 9.7 cm, removing reliability weighting increases it to 8.9 cm, replacing FedProx with FedAvg increases it to 10.9 cm, and removing uncertainty calibration raises the expected calibration error from 0.036 to 0.082. Figure 5(b) compares normalized accuracy, calibration, latency, communication efficiency, and robustness; the proposed model scores 0.91, 0.88, 0.84, 0.86, and 0.89, respectively. Figure 5(c) shows MAE, latency and communication volume for the five candidate models. The proposed model has an excellent region, with a 5.9cm MAE and 86ms client inference latency, and is 18.7MB per round; FedAvg-TinyViT uses 18.1MB but has an 8.1cm MAE, while centralized TinyViT reaches a 5.6cm MAE but requires raw data movement and is thus not a feasible distributed baseline. Figure 5(d) is the client drift measured by the norm of local update deviation. Median drift is reduced from 0.184 under FedAvg to 0.113 under FedProx-TinyViT, and the 90th percentile decreases from 0.332 to 0.201. Previous federated optimization work has determined that this reduction is due to the persistence of local objectives during aggregation [24].

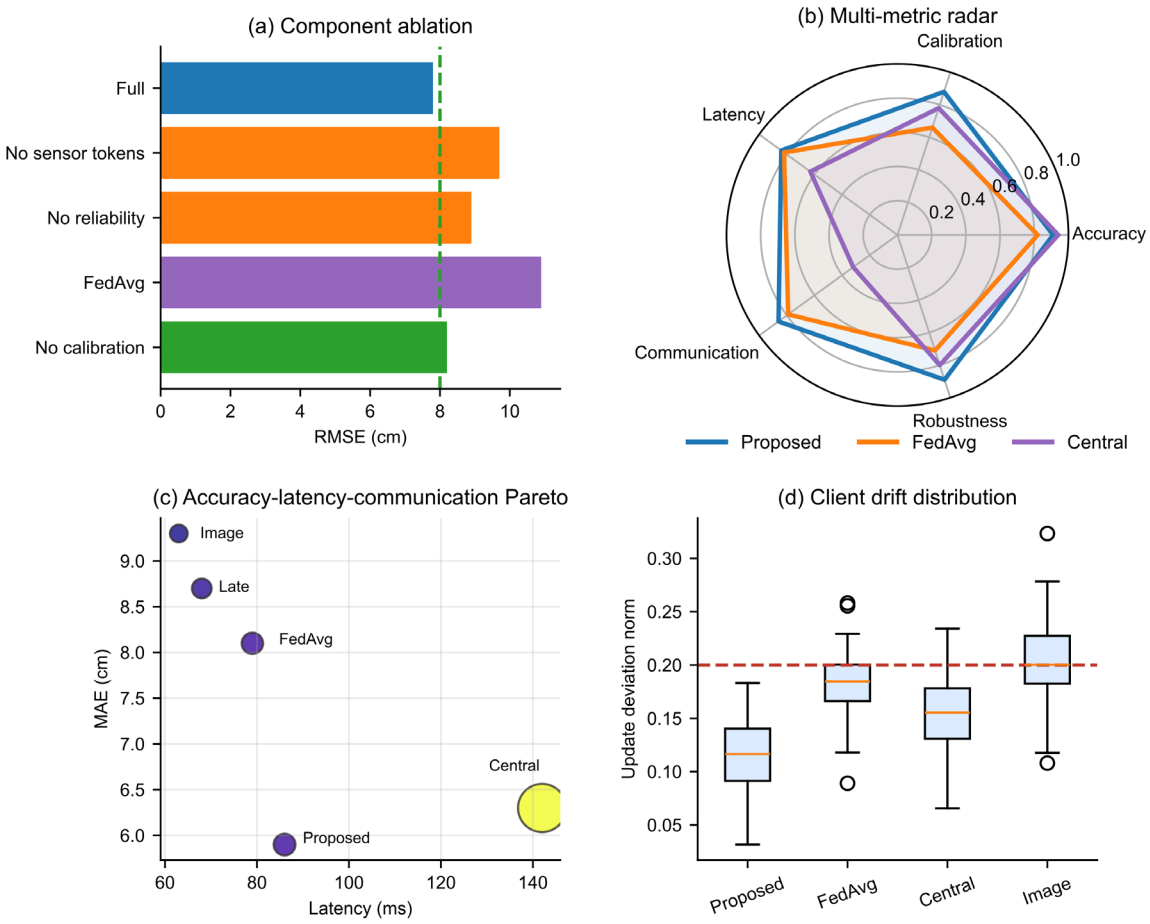

Figure 5. Ablation and efficiency analysis: (a) component-wise error change; (b) multi-metric radar comparison; (c) accuracy-latency-communication Pareto profile; (d) client drift distribution.

Robustness, Spatial Reconstruction, and Operational Alerting

Robustness tests simulate missing sensor readings, blurred nighttime images, and delayed crowdsourced uploads. Figure 6 is used at this point because robustness has to be read both as a missing-rate curve and as a spatially uneven district effect. Figure 6(a) reports MAE under missing sensor rates from 0% to 50%. The proposed model increases from 5.9 cm to 8.7 cm, whereas late fusion increases from 7.6 cm to 12.9 cm and sensor-only modeling fails rapidly after 30% missingness. Figure 6(b) provides a district-by-modality error heat map; the highest cell is 14.8 cm for image-only inference in the industrial underpass district, while the proposed model remains below 9.6 cm in all six districts. Recent work on urban sensing has noted that sensor outages and

uneven citizen reports tend to cluster spatially rather than appear at random [25]. The heat-map pattern follows this behavior because low-lying industrial roads contain deeper water, fewer facade references, and more saturated probes.

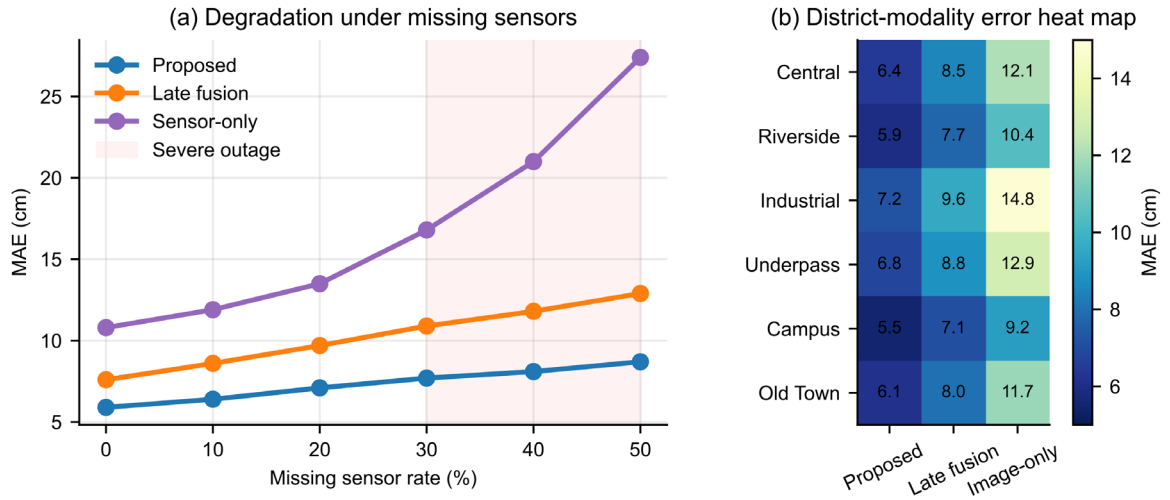

Figure 6. Robustness under missing or degraded observations: (a) performance degradation under missing sensor rates; (b) district-by-modality error heat map.

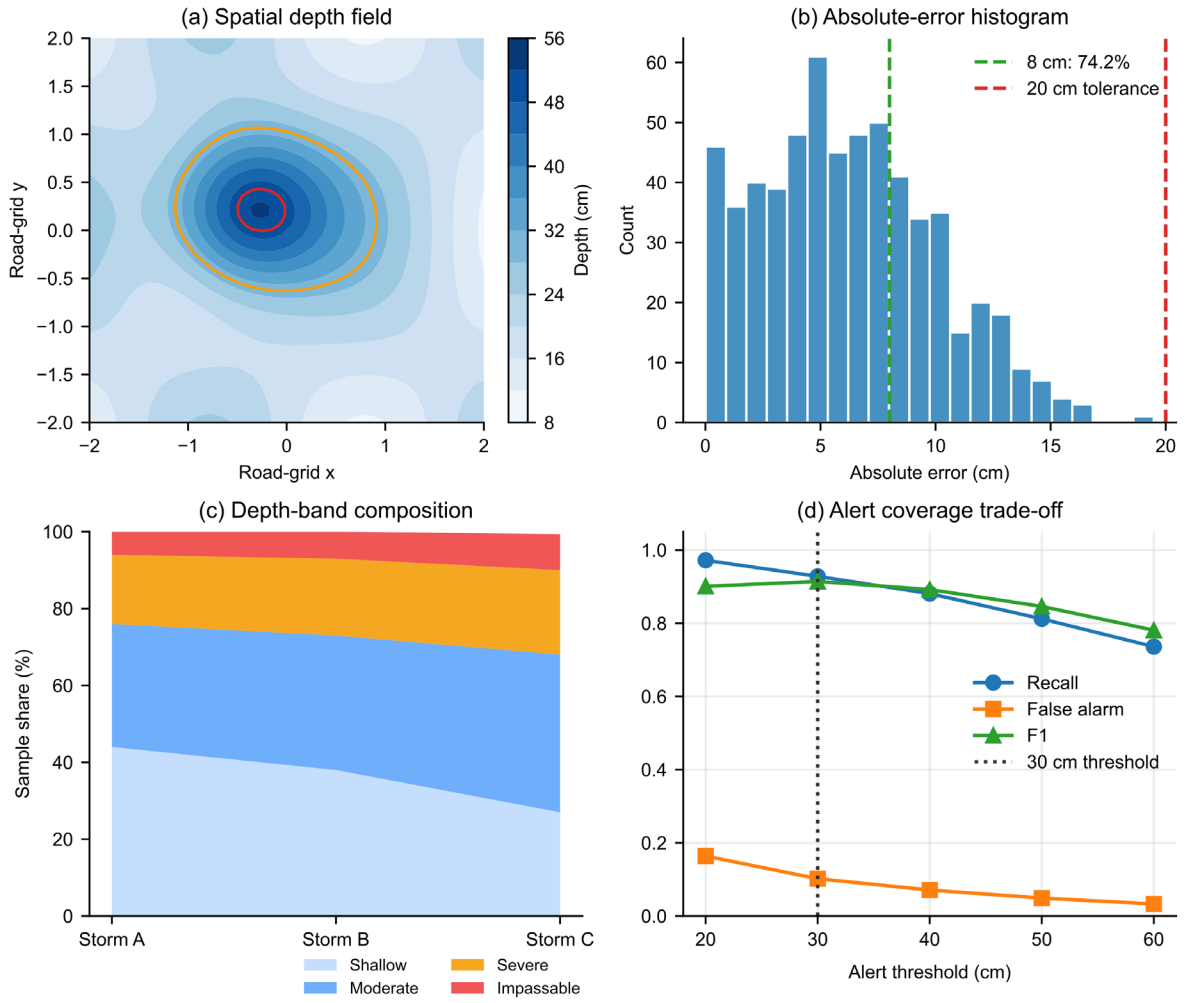

Figure 7. Operational reconstruction behavior: (a) spatial depth field on a flooded road grid; (b) absolute-error histogram with alert tolerance; (c) depth-band composition across events; (d) alert coverage and false-alarm trade-off.

Finally, analyze the reconstructed flood surface and warning behavior. Figure 7 is shown after the robustness results because it links the quality of numerical reconstruction with the map and threshold products that operators will see. Figure 7(a) shows a 40x40 road-grid depth field, and the proposed model has preserved a 58cm low-point basin and a 23cm curbside gradient near the main arterial. As shown in Figure 7(b), 74.2% of the absolute errors are less than 8cm, and only 5.8% exceed the 20cm tolerance for high-risk routing decisions. Figure 7(c) divides event composition into shallow, moderate, severe and impassable bands; the third storm has 31.4% severe or impassable samples and is therefore a more difficult test event. Figure 7(d) shows the alert coverage and false-alarm rate at different thresholds of 20cm - 60cm. At the 30 cm operational threshold, the proposed model obtains 0.914 alert F1, 0.928 recall, and 0.102 false-alarm rate, which is more balanced than image-only inference under low illumination [26]. Calibration is still necessary to differentiate between high-confidence shallow estimates and uncertain near-threshold cases by the response team [27]. A drop in the communication volume of 43.6% compared to raw-data upload also supports federated deployment under conditions of privacy and bandwidth restrictions [28].

A particular division of the area will be specified. The central commercial area has a relatively high density of building edges and curb references; the proposed model achieved a 5.6cm MAE and was close to that of centralized TinyViT. In the riverside area, there are sudden fluctuations in rainfall intensity and several gauges are near the peak of the event; therefore, reliability weighting reduces the effect of inconsistent probe readings and lowers the MAE from 8.4cm to 6.1cm. The Industrial Underpass District is a typical case. Inference based solely on images would underestimate deep water, as vehicle tires and curbs are often hidden. The proposed model uses sensor cues to keep the median prediction close to the surveyed depth range. The old-town district has narrow streets and strong night reflections, thus producing the largest uncertainty intervals. These intervals are not regarded as failures because they correctly identify cases with weak visual geometry. This behavior is more suitable for response planning; therefore, uncertain near-threshold sections will be routed to manual inspection instead of being silently classified as safe.

Training dynamics also show the effect of proximal regularization. In the first 20 communication rounds, FedAvg improves quickly for clients with dense images but oscillates for clients dominated by sensor readings. The proposed method converges more slowly during the first five rounds, then stabilizes after the model learns the shared waterline and sensor-prompt structure. By round 60, validation RMSE fluctuates within 0.4 cm for the proposed model, compared with 1.3 cm for FedAvg. The selected proximal coefficient of 0.01 provides a practical balance between local adaptation and aggregation stability.

The communication analysis uses serialized model updates rather than theoretical parameter counts. A full client update for the proposed backbone occupies 18.7 MB per round after 8-bit update quantization, and the optional reliability summary adds less than 0.2 MB. Raw-data upload for the same round would require approximately 33.2 MB for compressed images and sensor logs, even before privacy filtering. This comparison excludes video streams, which would make centralized collection substantially heavier. The training server therefore receives enough information to update the model while avoiding the most sensitive visual material. Edge latency remains under 100 ms on the tested municipal device profile, leaving time for local geocoding and dashboard update. The sensor prompt encoder contributes less than 4 ms, so most runtime cost still comes from visual patch processing.

Error inspection shows several repeated failure modes. First, reflective shop windows and wet asphalt may be mistaken for water surfaces when curbs or tire references are absent. Second, crowdsourced images captured from elevated buses can make shallow water appear deeper because the apparent waterline intersects the scene at a different angle. Third, a small number of gauges show delayed recession after peak rainfall, causing the sensor prompt to favor deeper estimates than the image alone. The reliability module reduces these cases but does not eliminate them. Therefore, uncertainty flags should be displayed together with point estimates in the operational dashboard.

Sensitivity tests of client participation show that the proposed framework remains feasible even when only half of the district clients join a round. With 12 of 24 clients sampled per round, RMSE increases from 7.8 cm to 8.3 cm. With 6 clients, RMSE reaches 9.2 cm and calibration error increases to 0.061. The degradation is moderate, indicating that reliability weighting helps prevent a small number of noisy participants from dominating aggregation. However, rare high-depth districts should be sampled more frequently during severe weather to capture the upper tail of the depth distribution.

Conclusion

This paper introduces an improved FedProx-TinyViT model for reconstructing urban flood depth based on crowdsourced images and distributed sensor data. The work framed depth reconstruction as a privacy-preserving federated multimodal regression problem, introduced sensor-conditioned prompt tokens into a lightweight transformer backbone, and used proximal optimization with reliability-weighted local evidence. The design of the model also considered the operational demands, such as missing observations, client drift, communication costs, uncertainty expression and road-grid reconstruction. The above method provides a feasible way for municipal-edge flood intelligence when raw image transmission is not feasible and client observations are heterogeneous. Compact Transformer models can be applied to hydrological emergency perception without a full-scale centralized data pipeline, as shown above.

Several limitations remain. The experimental benchmark approximates real municipal conditions but cannot cover all camera angles, drainage layouts, debris patterns and nighttime flood circumstances. The model also requires some high-quality depth labels for calibration. Severe weather that damages sensors, highly reflective water surfaces, or limited public participation may still cause bias in estimated depths around deep underpasses and complex road junctions.

Future research will link the model to physics-aware hydraulic constraints, a live traffic routing system and adaptive client selection strategy. Field deployment should also consider adding a human verification loop, privacy-preserving geolocation masking, and post-event active learning to improve the reconstruction system after each storm while maintaining accountability to municipal operators and the affected community. A feasible follow-up is to use federated updates, combine them with event-driven quality evaluations, and identify which districts need extra sensing, manual depth measurement or targeted camera deployment before the next communication cycle for emergency teams.

Author Contributions

Chen-fan Chien contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Kuang-yu Chuang contributes to analysis, investigation, data collection, draft preparation. Kuan-ming Liu contributes to data collection, draft preparation, manuscript editing. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Merz, B., Blöschl, G., Vorogushyn, S., Dottori, F., Aerts, J. C., Bates, P., ... & Macdonald, E. (2021). Causes, impacts and patterns of disastrous river floods. Nature Reviews Earth & Environment, 2(9), 592-609. https://doi.org/10.1038/s43017-021-00195-3

[2] Hofmann, J., & Schuettrumpf, H. (2021). floodGAN: Using deep adversarial learning to predict pluvial flooding in real time. Water, 13(16), 2255. https://doi.org/10.3390/w13162255

[3] Bentivoglio, R., Isufi, E., Jonkman, S. N., & Taormina, R. (2022). Deep learning methods for flood mapping: A review of existing applications and future research directions. Hydrology and Earth System Sciences, 26(16), 4345-4378. https://doi.org/10.5194/hess-26-4345-2022

[4] Cao, Y., Zhu, X., Liu, B., & Nan, Y. (2020). A qualitative study of the critical conditions for the initiation of mine waste debris flows. Water, 12(6), 1536. https://doi.org/10.3390/w12061536

[5] Ran, Q., Miao, S., Ye, Z., Tang, C., & Xu, J. (2022). Prediction of urban flood inundation using temporal-spatial deep learning. Water, 14(5), 722. https://doi.org/10.3390/w14050722

[6] Tellman, B., Sullivan, J. A., Kuhn, C., Kettner, A. J., Doyle, C. S., Brakenridge, G. R., Erickson, T. A., & Slayback, D. A. (2021). Satellite imaging reveals increased proportion of population exposed to floods. Nature, 596(7870), 80-86. https://doi.org/10.1038/s41586-021-03695-w

[7] Bates, P. D., Quinn, N., Sampson, C., Smith, A., Wing, O., Sosa, J., Savage, J., Olcese, G., Neal, J., Schumann, G., Giustarini, L., Coxon, G., Porter, J., Amodeo, M. F., Chu, Z., Lewis-Gruss, S., Freeman, N. B., Houser, T., Delgado, M., ... Krajewski, W. F. (2021). Combined modeling of US fluvial, pluvial, and coastal flood hazard under current and future climates. Water Resources Research, 57(2), e2020WR028673. https://doi.org/10.1029/2020WR028673

[8] Wing, O. E. J., Lehman, W., Bates, P. D., Sampson, C. C., Quinn, N., Smith, A. M., Neal, J. C., Porter, J. R., & Kousky, C. (2022). Inequitable patterns of US flood risk in the Anthropocene. Nature Climate Change, 12(2), 156-162. https://doi.org/10.1038/s41558-021-01265-6

[9] Sit, M., Demir, I., & Xiang, Z. (2020). A comprehensive review of deep learning applications in hydrology and water resources. Water Science and Technology, 82(12), 2635-2670. https://doi.org/10.2166/wst.2020.369

[10] Abdelkader, M., Shaqura, M., Claudel, C. G., & Gueaieb, W. (2020). A UAV based system for real time flash flood monitoring in desert environments using Lagrangian microsensors. Journal of Flood Risk Management, 13(1), e12553. https://doi.org/10.1111/jfr3.12553

[11] Jiang, S., Qiu, X., & Cao, C. (2021). Mapping urban flooding from social media and remote sensing observations. Remote Sensing, 13(10), 1930. https://doi.org/10.3390/rs13101930

[12] Leitao, J. P., Moy de Vitry, M., Scheidegger, A., & Rieckermann, J. (2021). Assessing the quality of digital elevation models obtained from mini unmanned aerial vehicles for overland flow modelling in urban areas. Hydrology and Earth System Sciences, 25(6), 3137-3156. https://doi.org/10.5194/hess-25-3137-2021

[13] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50-60. https://doi.org/10.1109/MSP.2020.2975749

[14] Lim, W. Y. B., Luong, N. C., Hoang, D. T., Jiao, Y., Liang, Y. C., Yang, Q., Niyato, D., & Miao, C. (2020). Federated learning in mobile edge networks: A comprehensive survey. IEEE Communications Surveys & Tutorials, 22(3), 2031-2063. https://doi.org/10.1109/COMST.2020.2986024

[15] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascon, A., Ghazi, B., Gibbons, P. B., ... Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1-2), 1-210. https://doi.org/10.1561/2200000083

[16] Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., & Gao, Y. (2021). A survey on federated learning. Knowledge-Based Systems, 216, 106775. https://doi.org/10.1016/j.knosys.2021.106775

[17] Nguyen, D. C., Ding, M., Pathirana, P. N., Seneviratne, A., Li, J., & Poor, H. V. (2021). Federated learning for Internet of Things: A comprehensive survey. IEEE Communications Surveys & Tutorials, 23(3), 1622-1658. https://doi.org/10.1109/COMST.2021.3075439

[18] Mothukuri, V., Khare, P., Parizi, R. M., Pouriyeh, S., Dehghantanha, A., & Srivastava, G. (2021). A survey on security and privacy of federated learning. Future Generation Computer Systems, 115, 619-640. https://doi.org/10.1016/j.future.2020.10.007

[19] Niknam, S., Dhillon, H. S., & Reed, J. H. (2020). Federated learning for wireless communications: Motivation, opportunities, and challenges. IEEE Communications Magazine, 58(6), 46-51. https://doi.org/10.1109/MCOM.001.1900461

[20] Liu, Y., James, J., Kang, J., Niyato, D., & Zhang, S. (2020). Privacy-preserving traffic flow prediction: A federated learning approach. IEEE Internet of Things Journal, 7(8), 7751-7763. https://doi.org/10.1109/JIOT.2020.2991401

[21] Wu, K., Zhang, J., Peng, H., Liu, M., Xiao, B., Fu, J., & Yuan, L. (2022). TinyViT: Fast pretraining distillation for small vision transformers. Lecture Notes in Computer Science, 13681, 68-85. https://doi.org/10.1007/978-3-031-20083-0_5

[22] Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., & Yan, S. (2022). MetaFormer is actually what you need for vision. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 10819-10829. https://doi.org/10.1109/CVPR52688.2022.01055

[23] Guo, J., Han, K., Wu, H., Xu, C., Tang, Y., Xu, C., & Wang, Y. (2022). CMT: Convolutional neural networks meet vision transformers. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 12175-12185. https://doi.org/10.1109/CVPR52688.2022.01186

[24] Chen, C. F. R., Fan, Q., & Panda, R. (2021). CrossViT: Cross-attention multi-scale vision transformer for image classification. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, 357-366. https://doi.org/10.1109/ICCV48922.2021.00042

[25] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, 10012-10022. https://doi.org/10.1109/ICCV48922.2021.00986

[26] Li, Y., Zhang, K., Cao, J., Timofte, R., & Van Gool, L. (2022). LocalViT: Bringing locality to vision transformers. International Journal of Computer Vision, 130(5), 1254-1274. https://doi.org/10.1007/s11263-021-01580-3

[27] Heo, B., Yun, S., Han, D., Chun, S., Choe, J., & Oh, S. J. (2021). Rethinking spatial dimensions of vision transformers. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, 11936-11945. https://doi.org/10.1109/ICCV48922.2021.01175

[28] Guo, M. H., Lu, C. Z., Liu, Z. N., Cheng, M. M., & Hu, S. M. (2022). Visual attention network. Computational Visual Media, 8(3), 331-346. https://doi.org/10.1007/s41095-022-0272-y