

Lightweight Cross-Modal Diffusion Networks for Real-Time Semantic Segmentation on Edge Devices

Jeremi Dryka

Department of Computer Science, University of New Hampshire, Durham, NH, USA.

jeremidryka535@unh.edu

Abstract

Real-time semantic segmentation on edge devices is a foundational capability for autonomous mobility, assistive robotics, augmented reality, industrial inspection, and disaster response. Although deep segmentation networks have achieved high accuracy, their computational intensity and memory demands conflict with the latency, energy, thermal, and bandwidth constraints of edge environments. Diffusion models have recently emerged as powerful generative architectures for dense prediction because they can iteratively refine uncertain regions and fuse heterogeneous sensing modalities. However, their iterative sampling procedures and large parameter counts make direct deployment difficult. This paper examines lightweight cross-modal diffusion networks as a system-level architecture for real-time semantic segmentation on edge devices. It analyzes how compact encoders, cross-modal fusion, latent diffusion decoders, distillation, quantization, and adaptive scheduling interact with hardware heterogeneity, infrastructure governance, sustainability, robustness, fairness, and policy. The discussion emphasizes structural trade-offs among accuracy, latency, energy, privacy, and maintainability rather than isolated model benchmarks. Case illustrations from mobility, medical edge, and disaster response show that cross-modal diffusion can improve resilience when one sensor degrades, but only if runtime orchestration, calibration, model updates, and accountability mechanisms are designed together. The paper concludes that lightweight diffusion segmentation should be treated as a socio-technical infrastructure problem, requiring reproducible evaluation, energy accounting, fairness auditing, and deployment governance.

Keywords

edge computing; semantic segmentation; diffusion models; cross-modal fusion; lightweight neural networks; real-time inference; socio-technical systems; sustainability; fairness.

1. Introduction

Semantic segmentation is one of the central perception tasks in modern intelligent systems because it transforms raw sensor streams into spatially organized meaning. In autonomous vehicles, robots, medical imaging devices, and augmented reality platforms, segmentation supports obstacle detection, drivable surface estimation, anatomical boundary identification, and scene understanding. The deep learning revolution described by LeCun, Bengio, and Hinton [1] established representation learning as a dominant paradigm, and fully convolutional networks demonstrated that dense prediction could be trained end to end [2]. Architectures such as U-Net and DeepLab further improved boundary precision and contextual aggregation through encoder-decoder structures and atrous convolution [3], [4]. These advances created the technical foundation for contemporary segmentation systems, but they also introduced substantial computational burdens that are difficult to reconcile with edge deployment.

The first response to this burden was the development of efficient neural architectures. MobileNet variants introduced depthwise separable convolutions and inverted residual bottlenecks to reduce multiply-accumulate operations and memory traffic [5], [6]. EfficientNet pursued compound scaling to balance depth, width, and resolution under resource constraints [7]. These backbones made real-time inference more plausible on mobile and embedded platforms, yet segmentation remains more demanding than classification because it requires high-resolution feature preservation and dense per-pixel outputs. When segmentation models are combined with multi-modal sensing, the resource profile grows further because depth, thermal, infrared, event-based, and language-conditioned inputs must be synchronized, encoded, and fused. The resulting system is not merely a neural network but a distributed sensing and computation pipeline whose behavior depends on hardware, software, environmental conditions, and governance choices.

Diffusion models offer a different route to robust dense prediction. Denoising diffusion probabilistic models learn to reverse a gradual noise process and have achieved strong generative fidelity [8]. Latent diffusion reduces the computational cost by operating in a compressed latent space [10], while improved sampling and architecture choices have narrowed the quality gap between diffusion and adversarial methods [11]. For segmentation, diffusion can be framed as iterative mask refinement, where uncertain regions are progressively clarified under the guidance of image features. Segmentation-oriented diffusion methods such as SegDiff have shown that probabilistic refinement can capture ambiguity and spatial uncertainty [12]. A recent diffusion-driven semantic segmentation approach with multi-modal image fusion also indicates that cross-modal guidance can enhance perception in complex scenes [9]. However, the iterative nature of diffusion conflicts with the strict latency budgets of real-time edge systems, creating a need for lightweight, adaptive, and cross-modal designs.

The system-level problem is therefore broader than model compression. Edge computing research has long emphasized that computation should be placed near data sources to reduce latency, preserve privacy, and improve resilience when connectivity is unreliable [18], [19], [20]. Model compression and knowledge distillation provide techniques for reducing parameter counts and inference costs [16], [17]. Cross-modal transformers and vision-language pretraining expand the representational capacity available to perception systems [13], [14], [15]. Yet these techniques must be integrated into an operational architecture that accounts for thermal throttling, battery depletion, sensor calibration drift, network outages, model updates, and accountability. This paper argues that lightweight cross-modal diffusion networks for edge segmentation should be understood as socio-technical infrastructures. The following sections examine their conceptual foundations, architectural design, deployment trade-offs, evaluation requirements, and policy implications.

2. Related Work and Conceptual Foundations

The evolution of semantic segmentation reflects a continuing tension between spatial precision and computational efficiency. Fully convolutional networks replaced fully connected classification heads with convolutional decoders, enabling dense prediction at arbitrary resolutions [2]. U-Net introduced skip connections that preserve fine-grained spatial information while recovering contextual semantics [3]. DeepLab combined atrous convolution with conditional random fields and later with encoder-decoder refinement to improve boundary adherence [4]. These architectures established the importance of multi-scale context, but they also demonstrated that accuracy often scales with receptive field size,

feature resolution, and model depth. For edge devices, those factors directly affect memory bandwidth, cache behavior, and energy consumption.

Efficient backbone design addresses part of this tension. MobileNetV2 uses inverted residuals and linear bottlenecks to reduce computation while maintaining representational capacity [5]. MobileNets provided an early blueprint for mobile vision through width and resolution multipliers [6]. EfficientNet systematically scaled network dimensions and achieved strong accuracy-efficiency trade-offs [7]. Nevertheless, segmentation heads, attention modules, and multi-modal fusion layers can dominate runtime even when the backbone is compact. This is especially true when cross-modal attention is used to align visual, depth, thermal, and textual representations. Vision transformers and Swin transformers offer powerful global context modeling [14], [15], but their memory access patterns can be poorly suited to embedded accelerators unless windowing, token pruning, or hybrid convolutional designs are used.

Diffusion models introduce a probabilistic refinement process that differs fundamentally from feed-forward segmentation. Denoising diffusion probabilistic models learn a sequence of denoising steps that gradually transform noise into structured data [8]. Latent diffusion models reduce the cost of this process by moving it into a learned latent space [10]. Improved sampling and classifier guidance have increased sample quality and controllability [11]. For segmentation, the diffusion process can be conditioned on image features and multi-modal cues, allowing the model to resolve ambiguous boundaries through iterative refinement [12]. A recent cross-modal diffusion segmentation approach demonstrates that multi-modal image fusion can enhance perception when a single modality is insufficient [9]. The core challenge is that iterative refinement multiplies inference cost by the number of sampling steps, which is often incompatible with real-time edge constraints.

Cross-modal learning provides complementary information that can improve robustness. Vision-language pretraining such as CLIP aligns images and text, enabling zero-shot and prompt-conditioned perception [13]. Transformer architectures provide flexible attention mechanisms for relating distant regions and heterogeneous tokens [14], [15]. In edge segmentation, cross-modal fusion can combine RGB with depth for geometry, thermal for low-light conditions, event streams for high-speed motion, and language prompts for task specification. However, fusion also introduces synchronization errors, calibration drift, missing modalities, and increased bandwidth. The required fusion must therefore be lightweight and adaptive, degrading gracefully when one modality is unavailable. Edge computing and edge intelligence research provide the infrastructure context for these designs [18], [19], [20]. Model compression and distillation offer methods to reduce the cost of both encoders and diffusion decoders [16], [17]. Sustainability, fairness, and maintainability concerns add further constraints that cannot be treated as afterthoughts [21], [22], [23], [24], [25]. The gap addressed in this paper is the integration of these concerns into a coherent architecture for real-time edge segmentation.

3. Lightweight Cross-Modal Diffusion Architecture for Edge Segmentation

A lightweight cross-modal diffusion network for edge segmentation can be understood as a layered system. The sensing layer acquires and time-aligns RGB, depth, thermal, event, inertial, and possibly textual inputs. The encoding layer uses compact convolutional or hybrid transformer backbones to extract modality-specific features [5], [6], [7], [14], [15]. The fusion layer aligns and combines these features through lightweight cross-attention, gated fusion, or token pruning. The diffusion layer generates or refines segmentation masks through a small number of denoising steps, ideally in a latent space to reduce spatial resolution and memory

traffic [10], [12]. The runtime layer schedules computation across CPU, GPU, neural processing unit, and occasionally cloud or nearby edge servers. Each layer introduces trade-offs that shape accuracy, latency, energy, and robustness.

Cross-modal fusion is central to the architectural promise. A depth sensor can disambiguate scale and support geometry-aware segmentation when visual texture is weak. A thermal camera can preserve perception in darkness, smoke, or glare. Event sensors can provide low-latency motion cues without exposing identifiable imagery. Text prompts can guide segmentation toward task-relevant categories, reducing the need for multiple specialized models [13]. However, fusion must avoid simply concatenating large feature maps, because that increases memory bandwidth and energy. Lightweight fusion can use modality-specific bottlenecks, dynamic gating, and cross-attention restricted to low-resolution tokens. When a modality is missing or unreliable, the system should reweight the remaining inputs rather than fail abruptly. This resilience is a system property, not merely a model property.

Diffusion refinement offers advantages for ambiguous boundaries. Instead of producing a single deterministic mask, a diffusion decoder can model uncertainty and iteratively correct errors. In edge contexts, however, the number of denoising steps must be aggressively reduced. Distillation from a larger diffusion teacher can compress the refinement process into fewer steps [17]. Quantization and pruning can reduce the cost of each step [16]. Latent diffusion can lower resolution while preserving semantic structure [10]. Segmentation-oriented diffusion can condition each step on multi-scale features and cross-modal cues [12]. A recent cross-modal diffusion design shows that guided fusion can improve perception under challenging conditions [9]. The architectural challenge is to preserve enough iterative refinement to improve robustness while meeting hard deadlines. Adaptive step scheduling, early exiting, and confidence-based stopping can help, but they require runtime monitoring and fallback strategies.

The deployment architecture must also account for model partitioning. Some layers may run on a neural processing unit, while diffusion refinement may run on a GPU or digital signal processor. If the edge device cannot meet the deadline, computation may be offloaded to a nearby server or cloud instance. Offloading can improve accuracy but increases latency variance, bandwidth cost, and privacy risk. Edge computing research emphasizes that placement decisions should be dynamic and context-aware [18], [19], [20]. Federated learning can update models across devices without centralizing raw data, but it introduces communication, heterogeneity, and security challenges [24]. Technical debt in machine learning systems can accumulate when data dependencies, configuration, and monitoring are poorly managed [25]. A sustainable architecture therefore requires observability, versioning, rollback, and clear ownership of model updates.

4. System-Level Trade-Offs in Real-Time Deployment

Real-time segmentation on edge devices is governed by latency budgets that vary by application. Autonomous driving may require perception updates within tens of milliseconds, while industrial inspection and medical triage may tolerate somewhat longer intervals but demand high reliability. Diffusion refinement can consume a large fraction of the budget, especially when multiple denoising steps are used. The system must therefore balance accuracy, latency, jitter, throughput, and energy. Thermal throttling is a particularly important constraint because sustained inference can heat a mobile device and reduce clock speeds. Battery-powered devices must trade computation against operational time. These constraints

mean that the best model on a desktop benchmark may be unsuitable at the edge even if its parameter count is small.

The edge-cloud continuum provides one response to these constraints. Latency-sensitive and privacy-sensitive layers can run locally, while heavier diffusion refinement can be offloaded when connectivity is strong. However, offloading introduces new dependencies. Network outages, congestion, or policy restrictions can make the system unavailable. Privacy regulations may prohibit transmitting raw video or medical images. Federated learning and secure aggregation can reduce data exposure, but they do not eliminate inference-time privacy risks [24]. A robust architecture should support multiple operating modes: a local lightweight mode, a cross-modal mode that uses all available sensors, and a hybrid mode that offloads only non-sensitive refinement. The system should degrade gracefully rather than fail catastrophically.

Cross-modal sensing adds significant infrastructure complexity. Cameras, depth sensors, thermal imagers, and event cameras have different frame rates, fields of view, and calibration procedures. Time synchronization errors can misalign features and confuse fusion layers. Calibration drift over time can silently degrade segmentation. Missing modalities may occur because of hardware failure, occlusion, power saving, or privacy settings. The fusion mechanism must therefore be trained with modality dropout and evaluated under realistic sensor degradation. Robustness is not only about adversarial attacks; it also concerns dust, rain, fog, vibration, and low light. A system that performs well only under ideal conditions is not deployable in safety-critical environments.

Governance and maintainability are equally important. Edge devices are often geographically distributed and difficult to update. Model versioning must be coordinated with sensor firmware, runtime libraries, and hardware accelerators. Rollbacks must be possible when an update causes regressions. Observability must capture latency, energy, confidence, sensor health, and fairness indicators without violating privacy. Technical debt can accumulate when training data, feature pipelines, and deployment configurations are not documented [25]. Sustainability concerns arise from the energy consumed by continuous inference and from the embodied carbon of hardware. Energy and policy considerations for deep learning have shown that computational efficiency is not only a technical metric but also an environmental and governance issue [21]. A responsible deployment must measure energy per inference and account for lifecycle impacts.

Fairness is another system-level concern. Segmentation models can perform unevenly across skin tones, clothing, weather conditions, geographic contexts, and disability-related use cases. Bias and fairness research has documented how machine learning systems can reproduce and amplify structural inequities [23]. Large-scale models can also concentrate power and exclude communities that lack data, hardware, or connectivity [22]. In edge deployment, fairness is affected by sensor availability, calibration quality, and update frequency. A device in a well-resourced hospital may receive frequent updates and high-quality sensors, while a device in a rural or disaster-affected area may operate with degraded inputs. Lightweight diffusion models should therefore be evaluated not only by average accuracy but also by worst-group performance, calibration, and failure modes.

5. Evaluation, Robustness, and Case Illustrations

Evaluation of lightweight cross-modal diffusion networks must go beyond mean intersection over union. A useful evaluation suite should include boundary accuracy, uncertainty

calibration, latency percentiles, energy per inference, memory peak, thermal behavior, and robustness to missing modalities. Because diffusion is stochastic, variance across sampling runs matters. A model that occasionally produces excellent masks but sometimes fails may be unacceptable in safety-critical use. Evaluation should also include domain shift, such as moving from daytime urban scenes to nighttime rural scenes, from clear weather to fog, or from one camera sensor to another. Cross-modal fusion can improve robustness, but only if the system detects when a modality is unreliable and reweights it appropriately.

Case illustrations clarify these trade-offs. In autonomous mobility, a lightweight diffusion segmentation system might fuse RGB, depth, thermal, and event streams to detect pedestrians, vehicles, and drivable surfaces. During daytime, RGB and depth may dominate; at night or in glare, thermal and event streams become more important. If the thermal sensor fails, the system should reduce confidence and increase safety margins rather than produce overconfident masks. In assistive robotics, segmentation can support navigation for visually impaired users or manipulation for wheelchair-mounted arms. Latency and privacy are paramount, and textual prompts may allow personalized object categories without retraining. In disaster response, drones and ground robots may operate with intermittent connectivity and damaged sensors. A lightweight diffusion model that runs locally and fuses whatever modalities remain available can support search and rescue, but it must be transparent about uncertainty.

Medical edge provides another important case. Portable ultrasound, endoscopy, and dermatology devices can benefit from real-time segmentation, but regulatory constraints, patient privacy, and clinical validation requirements are stringent. A diffusion model that refines ambiguous lesions or anatomical boundaries may improve interpretation, but stochastic outputs complicate accountability. Clinicians need explanations, confidence intervals, and audit trails. A cross-modal approach might fuse ultrasound with optical or thermal data, but each modality introduces calibration and consent issues. The system must therefore be designed with human oversight, logging, and clear escalation pathways. Similar concerns apply to industrial inspection, where false negatives can cause safety failures and false positives can halt production.

Robustness evaluation should include adversarial and non-adversarial threats. Adversarial patches, sensor spoofing, and data poisoning can distort segmentation. Non-adversarial threats include rain, dust, vibration, and power fluctuations. Cross-modal diffusion may improve robustness by using redundant cues, but it can also create new attack surfaces if fusion weights are manipulated. Runtime monitoring should detect anomalies in sensor statistics, feature distributions, and output confidence. When anomalies are detected, the system should switch to a conservative mode, request human intervention, or reduce vehicle speed. These responses are governance decisions as much as technical ones. A responsible evaluation framework should therefore combine model metrics with systems metrics and organizational readiness.

6. Policy Implications and Forward-Looking Perspectives

Policy and governance must keep pace with the deployment of lightweight cross-modal diffusion networks. Safety-critical applications require certification processes that account for probabilistic models, iterative inference, and multi-modal fusion. Regulators may need to examine training data provenance, update procedures, failure modes, and human oversight. Transparency is difficult when diffusion models are large and stochastic, but it is not impossible. Model cards, datasheets, and audit logs can document intended use, limitations,

and performance across subgroups. Standards for edge artificial intelligence should address latency, energy, privacy, and interoperability. Without such standards, vendors may optimize for benchmarks while neglecting field reliability.

Data governance is equally important. Cross-modal sensing can capture intimate details about individuals and environments. Privacy-preserving techniques such as federated learning, differential privacy, and on-device processing can reduce data exposure [24]. However, these techniques have trade-offs in accuracy, communication, and complexity. Policy should clarify consent, retention, and secondary use. It should also address cross-border data flows and surveillance risks. Fairness audits should be required for public-sector deployments and for systems that affect access to mobility, healthcare, or employment. Bias research shows that technical fixes alone are insufficient; participatory design, community oversight, and redress mechanisms are necessary [23]. Large-scale models can concentrate economic and political power, so procurement policies should support smaller organizations, public interest research, and open benchmarks [22].

Sustainability policy should require energy reporting for edge artificial intelligence. The energy and carbon costs of training and inference are significant, and edge devices multiply the number of inference sites [21]. Lightweight diffusion can reduce per-inference cost, but continuous operation across millions of devices can still produce substantial aggregate energy demand. Hardware lifecycle, e-waste, and rare-earth mineral extraction must be considered alongside model efficiency. A circular economy approach would prioritize repairability, long-term software support, and modular sensor upgrades. Technical debt in machine learning systems is also a sustainability issue because poorly maintained models require repeated retraining and redeployment [25]. Sustainable design should favor reusable components, clear interfaces, and robust monitoring.

Looking forward, several technical directions could improve the feasibility of lightweight cross-modal diffusion. Event-based sensors and neuromorphic hardware may reduce latency and energy for dynamic scenes. On-device continual learning could adapt to local conditions without central data collection, though it raises stability and security concerns. Distillation and quantization will continue to reduce model size, but their interaction with diffusion sampling must be studied carefully [16], [17]. Hybrid architectures that combine convolutional efficiency with transformer context may offer better trade-offs than either alone [14], [15]. Cross-modal foundation models may provide strong priors, but they must be compressed and aligned with edge constraints [13]. The most promising path is not a single model but an adaptive system that selects modalities, sampling steps, and compute placement according to context.

7. Conclusion

Lightweight cross-modal diffusion networks offer a compelling direction for real-time semantic segmentation on edge devices because they combine the robustness of generative refinement with the complementary information of multiple sensing modalities. They can improve boundary accuracy, handle uncertainty, and degrade gracefully when sensors fail. However, their iterative inference and multi-modal fusion create significant system-level challenges. Latency, energy, thermal limits, synchronization, calibration, privacy, maintainability, and fairness must be addressed together. Treating these networks as isolated models obscures the infrastructure and governance decisions that determine whether they work in practice. The analysis in this paper suggests that deployment should be guided by adaptive scheduling, graceful degradation, energy accounting, fairness auditing, and clear

accountability. Policy frameworks should support transparency, privacy, sustainability, and equitable access. Future research should develop reproducible benchmarks that measure system performance across diverse hardware, environments, and populations. If designed responsibly, lightweight cross-modal diffusion segmentation can serve as a foundational perception capability for edge intelligence without reproducing the vulnerabilities and inequities of centralized, resource-intensive artificial intelligence.

References

1. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
2. Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3431–3440). IEEE.
3. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (pp. 234–241). Springer.
4. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4), 834–848.
5. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4510–4520). IEEE.
6. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv*. <https://doi.org/10.48550/arXiv.1704.04861>
7. Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 6105–6114). PMLR.
8. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems* (pp. 6840–6851). Curran Associates.
9. Wang Z, He D, Zhao L, et al. Difusionseg: Diffusion-driven semantic segmentation with multi-modal image fusion for enhanced perception[J]. *Knowledge-Based Systems*, 2025: 114481.
10. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10684–10695). IEEE.
11. Dhariwal, P., & Nichol, A. (2021). Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems* (pp. 8780–8794). Curran Associates.
12. Amit, T., Shaharbany, T., Nachmani, E., & Wolf, L. (2021). SegDiff: Image segmentation with diffusion probabilistic models. *arXiv*. <https://doi.org/10.48550/arXiv.2112.00390>

13. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 8748–8763). PMLR.
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.
15. 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 10012–10022). IEEE.
16. Han, S., Mao, H., & Dally, W. J. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. In *International Conference on Learning Representations*.
17. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv*. <https://doi.org/10.48550/arXiv.1503.02531>
18. Satyanarayanan, M. (2017). The emergence of edge computing. *Computer*, 50(1), 30–39.
19. Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. *IEEE Internet of Things Journal*, 3(5), 637–646.
20. Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., & Zhang, J. (2019). Edge intelligence: Paving the last mile of artificial intelligence with edge computing. *Proceedings of the IEEE*, 107(8), 1738–1762.
21. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3645–3650). Association for Computational Linguistics.
22. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). ACM.
23. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1–35.
24. 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., Gascón, 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.
25. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems* (pp. 2503–2511). Curran Associates.