

# Efficient and Privacy-preserving Smart Edge for Hierarchical Vision Systems

P. Pad, Y. Sepehri, N. Maamari

Edge-cloud systems are increasingly important in today's technology landscape, enabling devices with limited resources, like smartphones and IoT devices, to perform complex AI tasks by offloading some processes to the cloud. These systems allow real-time data processing, reduce latency, and help manage the constraints of memory and computation on edge devices. This work focuses on enhancing deep neural network training on such resource-limited devices by implementing a hierarchical edge-cloud framework that improves efficiency and privacy. A hierarchical training method divides training tasks between edge devices and cloud servers, using early exits to reduce communication and runtime. This approach keeps raw data on the edge device, reducing privacy risks and achieving a 60-80% reduction in training time with minimal impact on accuracy. To further protect sensitive information, adversarial early exits are combined with differential privacy techniques, preventing sensitive content from being inferred from transmitted feature maps while maintaining accuracy within 3%. Together, these methods enable secure and efficient deep learning on edge devices with support from the cloud.

Edge-cloud systems enable efficient processing for deep learning tasks on devices with limited resources by splitting computations between edge devices and cloud servers. This approach helps manage memory and computation constraints, latency, and data privacy. We introduce two key methods that by incorporating early exits, these methods reduce data communication and latency, enhance power efficiency, and safeguard privacy without compromising overall accuracy.

In the first method, an early exit is introduced at the edge device, allowing data to be processed locally without requiring transmitting the gradient from the cloud server to the edge during the backpropagation in the training phase. By reducing the need for data transmission, this approach significantly lowers communication latency and power consumption, making the training phase much more efficient. Since data transfers between edge and cloud consume considerable energy and time, minimizing these transmissions improves the overall training efficiency. Also, the edge in this method only needs to have the transmission capability and does not need to have a receiver which simplifies the edge device architecture. Additionally, only feature maps, rather than raw data, are sent to the cloud, enhancing privacy by reducing direct access to sensitive information. Without needing backpropagation gradients from the cloud, this setup also enables the use of non-differentiable processing techniques—like sophisticated compression algorithms—on the edge, further optimizing data handling and storage. Through extensive experiments, we showed that the proposed method reduces the training time 60%-80% while has a minimal impact on the overall inference accuracy<sup>[1]</sup>.

The second method expands on privacy protection by implementing an adversarial approach to the early exit at the edge. This method trains the edge device to remove sensitive information, such as personal attributes, from the data before it's sent to the cloud, and retaining only the task-relevant information. During training, the adversarial early exit learns to suppress sensitive content (e.g., a person's gender) while preserving features necessary for the task at hand. For instance, in an application to detect smiling in facial images, the model is trained to eliminate gender information, which cannot be recovered even with advanced processing on the cloud. This method strictly enhances privacy by ensuring that any potentially sensitive

information is filtered out before data leaves the edge, and only essential features for the specific task—such as a smile detection indicator—are transmitted. This setup ensures that data privacy is maintained throughout the training process, as any sensitive details are inaccessible to cloud-side operations.

Figure 1 shows some examples of the performance of the proposed method called PriPHiT against advanced reconstruction methods. We observe that if PriPHiT is not implemented, the reconstruction attack can very precisely reconstruct the input image from the extracted features of the neural network (e.g. VGG-11, ResNet-18 and MobileViT-xxs). However, by performing the training using the PriPHiT method, the gender of the object is mixed up in the reconstructed image which is even more powerful than preventing the image from being reconstructed<sup>[2]</sup>.

Desired Label Sensitive Label	Input	Deep Reconstruction Attack					
		VGG-11		ResNet-18		MobileViT-xxs	
		PriPHiT	Baseline	PriPHiT	Baseline	PriPHiT	Baseline
Month Open With Makeup							
Month Closed With Makeup							
Month Open No Makeup							
Month Closed No Makeup							

Figure 1: Examples of reconstruction of the input image from the features extracted with and without the PriPHiT method.

In conclusion, these two methods, both contribute to making edge-cloud deep learning systems more efficient and privacy-conscious. The first method emphasizes communication efficiency and system responsiveness by reducing latency and energy consumption, while the second ensures strict privacy by eliminating sensitive information at the edge. Together, they provide a robust framework for training deep learning models on edge devices while leveraging cloud capabilities in a secure and efficient manner.

[1] Sepehri, P. Pad, A. C. Yüzügüler, P. Frossard, L. A. Dunbar, Hierarchical Training of Deep Neural Networks Using Early Exiting, IEEE Transactions on Neural Networks and Learning Systems, doi: 10.1109/TNNLS.2024.3396628

[2] Y. Sepehri, P. Pad, P. Frossard, L. A. Dunbar, PriPHiT: Privacy-Preserving Hierarchical Training of Deep Neural Networks, arXiv:2408.05092