# Multi-View Images Over-the-Air Aggregation and Classification

## CSC2529 Project Report

Fan Yang

**Abstract**—Real-time processing of multi-view data in wireless sensing networks is challenging due to communication bottlenecks and limited channel resources. To address this, we propose a multi-view over-the-air aggregation framework that integrates split learning and over-the-air computation (AirComp). In this framework, the neural network is split into sensor-side feature extraction and server-side classification, enabling efficient transmission of compressed and aggregated feature representations instead of raw data. AirComp further enhances real-time performance by simultaneously aggregating signals over shared channels, significantly reducing communication overhead. Experimental results on the ModelNet40 dataset demonstrate the effectiveness of the proposed approach, achieving low latency and high classification accuracy under noisy conditions. This framework offers a scalable and robust solution for real-time processing in large-scale multi-view systems.

**Index Terms**—Multi-view Images, Machine Learning, Convolutional Neural Networks

## 1 Introduction

Multi-view learning is an advanced machine learning paradigm that jointly leverages data collected from multiple views or measurement methods. Each view offers a distinct set of features for the same underlying data instance, capturing complementary information to improve overall learning performance [1]. Multi-view learning has been widely adopted in diverse applications, ranging from computer vision and natural language processing to bioinformatics and wireless communication systems.

In recent years, the proliferation of modern multi-device sensing wireless networks has introduced new opportunities and challenges for multi-view classification tasks. These networks, equipped with various sensors and communication devices, collect heterogeneous data from different views or modalities, providing a rich foundation for advanced analytics. Multi-view classification in this context is considered a promising objective detection and recognition technology, offering the potential to improve system reliability, adaptability, and efficiency.

Traditional approaches to multi-view classification often involve deploying AI models either on edge devices or on a central server. While on-device inference avoids transmission delays, it imposes substantial computational overhead, especially when dealing with resource-intensive deep neural networks. Conversely, on-server inference alleviates computational constraints on devices but incurs significant communication overhead, as the high-dimensional raw data must be transmitted to the server. This trade-off between computation and communication represents a major bottleneck for practical deployment in wireless networks, especially under stringent latency and bandwidth constraints.

• *Fan Yang is with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada, M5S 1A1.*
*E-mail: fanchn.yang@mail.utoronto.ca*

To address these limitations, there is an urgent need for an efficient and scalable multi-view classification framework tailored to the characteristics of multi-device sensing wireless networks. This framework must minimize communication overhead while maintaining high inference accuracy and robustness, particularly in environments with noisy channels and limited resources.

In this paper, we tackle these challenges by proposing a novel multi-view over-the-air aggregation framework that synergizes split learning and over-the-air computation (AirComp). Our framework leverages the unique properties of AirComp to perform data aggregation directly within the wireless medium, significantly reducing communication costs while preserving data integrity. Furthermore, the integration of split learning enables efficient model training and inference across distributed devices without the need for transmitting raw data.

In summary, the main contributions of our work are as follows:

1) **A novel multi-view over-the-air aggregation framework:** We propose a hybrid approach that combines split learning and AirComp to address the communication bottlenecks in wireless sensing networks. This framework is designed to be scalable, efficient, and adaptable to various network conditions.
2) **Tunable aggregation for robust and efficient learning:** We introduce a flexible aggregation method that allows for adjustable trade-offs between noise robustness and pooling efficiency, enabling better performance under diverse channel conditions.
3) **Extensive evaluation on real-world datasets:** Our experiments demonstrate significant reductions in communication costs and high classification accuracy on the ModelNet40 dataset, even in the pres-

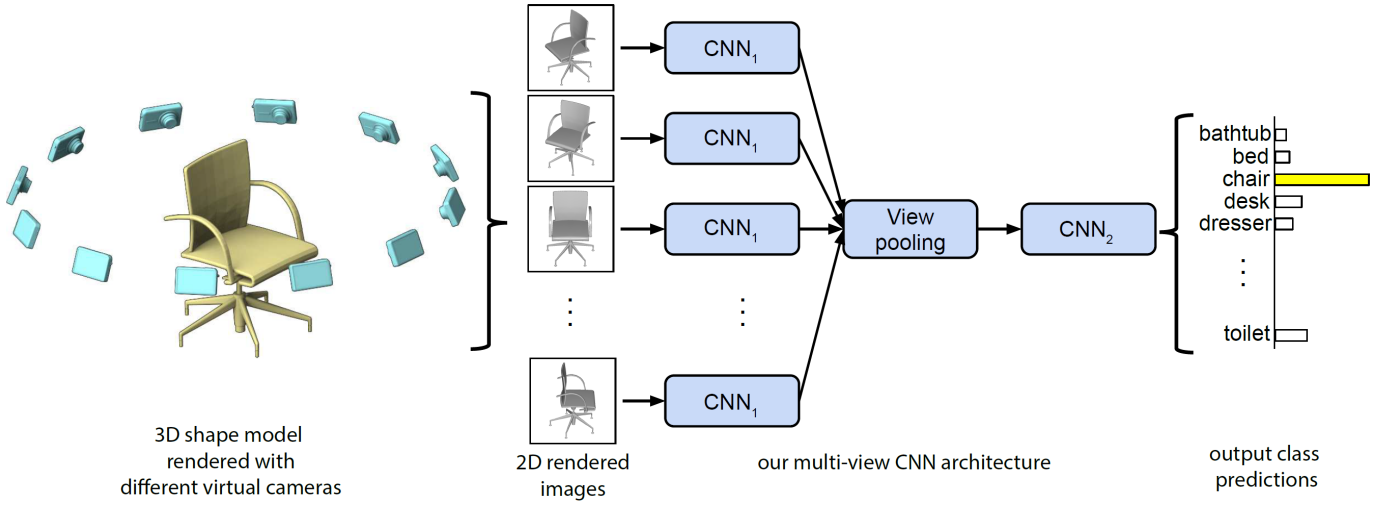


Fig. 1. Multi-view CNN for 3D shape recognition [1].

ence of noisy wireless channels. The results highlight the potential of our approach for practical deployment in multi-device sensing systems.

## 2 Related Works

### 2.1 Multi-view learning

The foundational concept of multi-view learning, known as co-training, was introduced by Blum and Mitchell [2]. Co-training involves training two classifiers on different views of the data to improve performance by leveraging complementary information. Xu et al. [3] later formalized the core principles of multi-view learning as consensus and complementary. The consensus principle seeks to maximize agreement between different views, while the complementary principle emphasizes leveraging unique knowledge provided by each view. Together, these principles enhance the generalization and robustness of multi-view models.

In recent years, the integration of multi-view learning with deep learning has gained traction due to the latter's ability to learn powerful representations. Pioneering works have applied deep learning techniques to multi-view tasks, such as multi-view canonical correlation analysis [4], multimodal feature extraction [5], and joint learning from multiple views [6]. These advancements enable more effective and scalable multi-view systems, laying the foundation for state-of-the-art applications in areas like 3D shape recognition, image classification, and multimodal sensor fusion.

### 2.2 AirComp

Over-the-air computation (AirComp) is an emerging technique that exploits the waveform-superposition property of multi-access wireless channels. This allows for simultaneous aggregation of features transmitted by multiple devices, enabling efficient data fusion without requiring dedicated communication channels [7]. By leveraging this property, AirComp reduces communication overhead and latency, making it particularly suitable for large-scale, resource-constrained systems. However, the approach faces challenges such as signal interference, noise amplification, and limited scalability under high device densities. Recent advancements focus on optimizing aggregation strategies and mitigating noise to enhance its robustness in practical deployment.

### 2.3 Split inference

Split inference addresses the computational and energy constraints of edge devices by dividing an AI model into two parts [8]. The sensor-side sub-model is deployed on resource-limited devices to perform lightweight feature extraction, while the server-side sub-model handles computationally intensive tasks, such as classification or prediction, at an edge server. This paradigm reduces the volume of data transmitted over the network by sending compressed feature representations instead of raw data. Recent developments in split inference explore adaptive partitioning strategies, dynamic task allocation, and optimization for heterogeneous network conditions. This approach is particularly effective for real-time applications, where latency and communication cost are critical considerations.

## 3 Proposed Method

To address the communication bottleneck problem in multi-view sensing wireless networks, we propose a multi-view over-the-air aggregation model, as illustrated in Figure 2. This model integrates split learning and AirComp to efficiently process multi-view data in a distributed manner, significantly reducing communication overhead while maintaining high classification accuracy.

### 3.1 Split Learning

In this system, the entire pretrained classification neural network is divided into two distinct parts: the sensor-side sub-model for feature extraction and the server-side sub-model for classification.

At the sensor side, $K$ sensors are responsible for capturing continuous images from different views of the same object. Each sensor processes its input image using the sensor-side sub-model, producing feature tensors $\{\mathbf{f}_k\} \in \mathbb{R}_+^N$. These

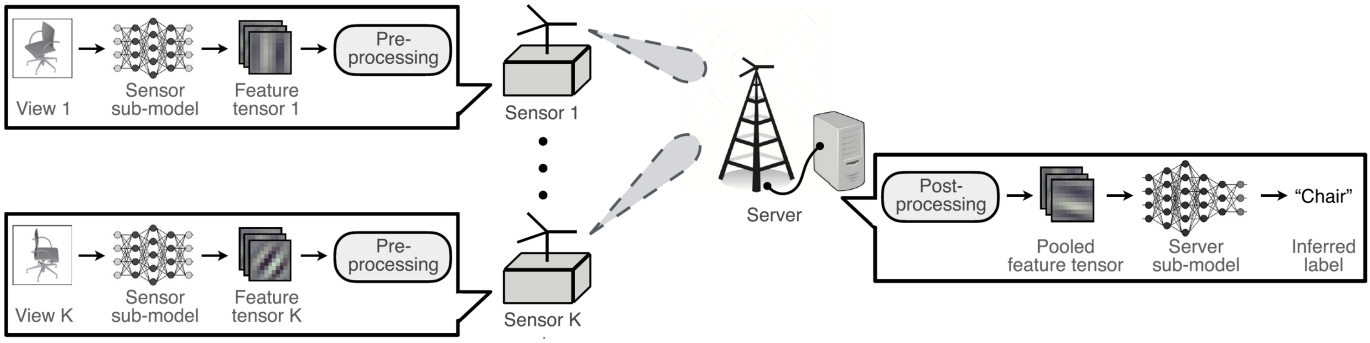


Fig. 2. Multi-view over-the-air aggregation and classification system model.

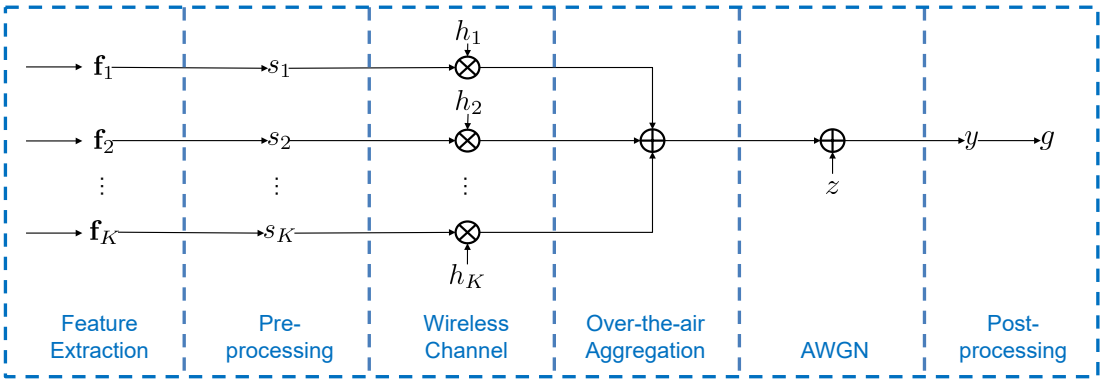


Fig. 3. System model for AirComp.

feature tensors, which are significantly lower in dimension compared to the original images, are pre-processed to generate transmitted signals $\{\mathbf{s}_k\}$. This compression drastically reduces the communication overhead while preserving the essential discriminative information of the data.

At the server side, the transmitted signals from all sensors are received as aggregated signals $\mathbf{y}$ through a shared wireless channel. After applying post-processing techniques, the server reconstructs the aggregated feature tensors $\mathbf{g}$, which are then fed into the server-side sub-model for classification. By avoiding the direct transmission of raw data, split learning ensures privacy preservation and computational efficiency at the sensor side.

### 3.2 AirComp

AirComp is a key enabler of the proposed framework, allowing signals from multiple sensors to be transmitted simultaneously over a single wireless channel, as illustrated in 3. Unlike conventional methods that require dedicated channels for each sensor, AirComp exploits the natural superposition property of wireless signals, resulting in significant communication efficiency gains.

During transmission, the signals from each sensor overlap, producing an aggregated signal directly in the wireless channel:

$$\mathbf{y} = \sum_{k=1}^{K} \mathbf{s}_k + \mathbf{z}, \tag{1}$$

where $\mathbf{z} \sim \mathcal{N}(0, \sigma^2)$ represents additive white Gaussian noise (AWGN).

This process effectively realizes over-the-air pooling during transmission, combining the contributions from all sensors into a single aggregated representation. Unlike traditional approaches that require a separate pooling layer at the server side, AirComp eliminates the need for explicit pooling operations by performing aggregation implicitly in the wireless medium. This feature not only reduces computational overhead at the server but also ensures efficient utilization of channel resources.

Although the overlap introduces some information loss due to noise and interference, the proposed framework is designed to tolerate and compensate for such distortions. The pre- and post-processing steps further enhance the quality of the aggregated features, ensuring that the classification performance remains robust even under noisy conditions.

### 3.3 Pre- and Post-processing

To implement a generalized over-the-air pooling method, we introduce a tunable exponential parameter $\alpha \geq 1$ to pre-process the features $\{\mathbf{f}_k\}$. Specifically, each element of transmitted signals at sensor $k$ is generated as

$$s_k = f_k^{\alpha}. \tag{2}$$

Correspondingly, each element of the aggregated features is calculated as

$$g = \frac{\left[(y)^{+}\right]^{\frac{1}{\alpha}}}{\beta} = \frac{\left[\left(\sum_{k=1}^{K} f_k^{\alpha} + z\right)^{+}\right]^{\frac{1}{\alpha}}}{\beta} \tag{3}$$

where the ramp function is defined as $(\cdot)^+ = \max\{\cdot, 0\}$. $\beta$ is a scaling factor to ensure that the aggregated feature $g$ retains the same scale as the original feature $\{f_k\}$.

Obviously, the value of $\beta$ depends on the selection of the exponential parameter $\alpha$. To calculate $\beta$, we assume a noise-free wireless channel and employ max-pooling as the reference, minimizing the pooling error:

$$\begin{aligned}\beta^* &= \arg\min_{\beta} \mathbb{E}\left[\left(\frac{1}{\beta}\left[\left(\sum_{k=1}^{K} f_k^{\alpha}\right)^{+}\right]^{\frac{1}{\alpha}} - f_{\max}\right)^2\right] \\ &= \arg\min_{\beta} \frac{1}{\beta^2}\mathbb{E}\left[\left[\left(\sum_{k=1}^{K} f_k^{\alpha}\right)^{+}\right]^{\frac{2}{\alpha}}\right] \\ &\quad - \frac{2}{\beta}\mathbb{E}\left[\left[\left(\sum_{k=1}^{K} f_k^{\alpha}\right)^{+}\right]^{\frac{1}{\alpha}} f_{\max}\right] + \mathbb{E}\left[f_{\max}^2\right] \\ &= \frac{\mathbb{E}\left[\left[\left(\sum_{k=1}^{K} f_k^{\alpha}\right)^{+}\right]^{\frac{2}{\alpha}}\right]}{\mathbb{E}\left[\left[\left(\sum_{k=1}^{K} f_k^{\alpha}\right)^{+}\right]^{\frac{1}{\alpha}} f_{\max}\right]}\end{aligned} \tag{4}$$

When channel noise is negligible, by tuning the exponential parameter $\alpha$, we can realize both average-pooling and max-pooling as follows:

$$\begin{cases}\alpha = 1, & \Longrightarrow \text{average-pooling} \\ \alpha \to \infty, & \Longrightarrow \text{max-pooling}\end{cases} \tag{5}$$

However, when the channel noise is non-negligible, max-pooling becomes more susceptible to noise than average-pooling. This is because max-pooling amplifies the contribution of the largest feature values, which are more likely to be significantly distorted by noise. In contrast, average-pooling aggregates information from all features, making it inherently more robust to noise. Consequently, there is a trade-off between max-pooling and average-pooling, which can be controlled by adjusting the parameter $\alpha$, as discussed in the next section.

## 4 Experiment

We evaluate the proposed method using the ModelNet40 dataset [9] and the ResNet18 model for training and testing within an MVCNN architecture. The total number of sensors is set to $K = 12$. The ResNet18 model is first trained without considering the impact of wireless channels and then split into two parts before the linear classifier. The classifier is deployed at the server side, while the remaining components are deployed at each sensor for feature extraction. As a result, the number of transmitted features at each sensor is $N = 512$.

In the following, we will present the classification performance of the proposed method.

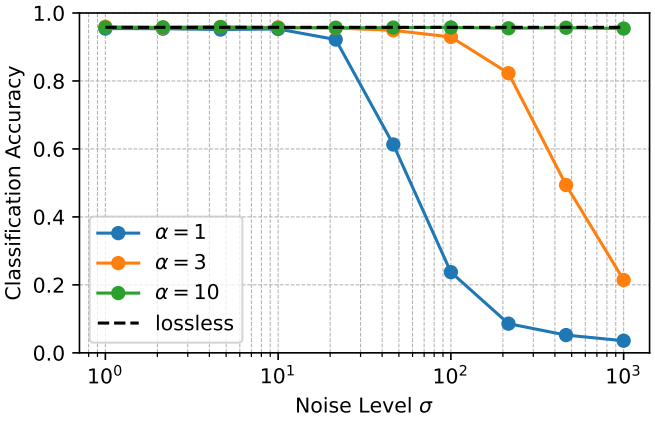


Fig. 4. Classification accuracy versus noise level $\sigma$.

### 4.1 Classification Accuracy vs. Noise

Figure 4 shows the classification accuracy as a function of the noise level $\sigma$. It can be observed that as the noise level increases, classification performance is indeed degraded. However, by adjusting the exponential parameter $\alpha$, the proposed method can effectively enhance classification performance under high noise conditions, achieving results that are even close to the classification accuracy of lossless transmission. This demonstrates the robustness and adaptability of the proposed method.

### 4.2 Classification Accuracy vs. Exponential Parameter

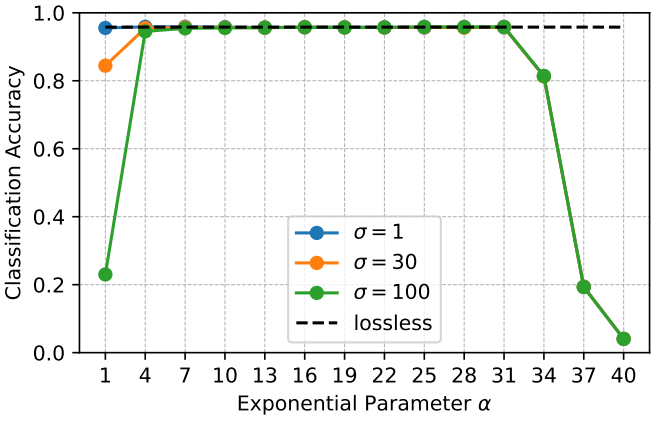


Fig. 5. Classification accuracy versus exponential parameter $\alpha$.

Figure 5 shows the classification accuracy as a function of the exponential parameter $\alpha$. It can be observed that increasing $\alpha$ under high noise conditions effectively enhances classification performance. However, as $\alpha$ continues to increase, the classification accuracy drops sharply. This decline occurs because larger $\alpha$ values approach max-pooling, which is more sensitive to noise due to its reliance on the largest feature values. This demonstrates the trade-off between average-pooling and max-pooling in the context of noisy transmission.

### 4.3 Communication Latency

We consider a wireless channel with a communication bandwidth of $B = 20$ MHz, which is a typical bandwidth for Wi-Fi systems. According to the Shannon theorem, the channel capacity, i.e., the maximum achievable lossless transmission rate, can be expressed as:

$$C = B \log_2(1 + \frac{P}{\sigma^2}) \tag{6}$$

where $P$ represents the signal power and $\sigma^2$ denotes the noise power. Here, the signal power $P$ has been normalized to simplify the expression.

The minimum communication latency for traditional lossless transmission can then be calculated as:

$$L_{\text{lossless}} = \frac{\text{image size} \times \text{number of sensors}}{\text{channel capacity}} \tag{7}$$

where the size of each image in the ModelNet40 dataset is approximately 12 kB.

In contrast, for our proposed method, since the sensors transmit signals over the same channel simultaneously, the required communication overhead for each classification task is the number of transmitted features per view, i.e., 512. Therefore, the communication latency of our proposed AirComp method is given by:

$$L_{\text{AirComp}} = \frac{\text{feature size}}{\text{bandwidth}} \tag{8}$$

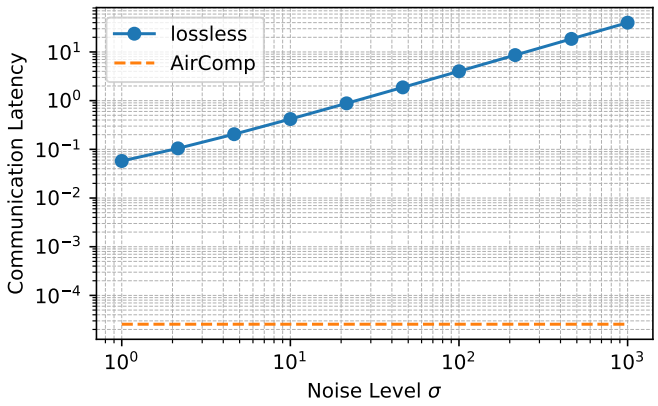


Fig. 6. Communication latency versus noise level $\sigma$.

Figure 6 illustrates the communication latency as a function of the noise level $\sigma$. It can be observed that the communication time required for traditional lossless transmission is significantly higher than that of our proposed method, which is based on split learning and AirComp. Furthermore, as the noise level increases, the communication overhead for traditional lossless transmission also grows, whereas the communication overhead for the proposed method remains constant.

### 4.4 Classification Accuracy vs. Quantization Level

To reduce the communication overhead of lossless transmission, a common approach is to quantize each pixel of the image data to a specified bit level. However, quantization introduces distortion to the original images, which can adversely affect classification performance.

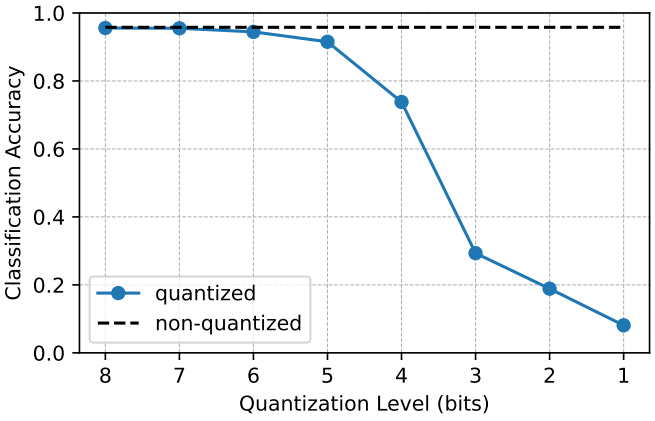


Fig. 7. Classification accuracy versus quantization level (bits).

Figure 7 shows the classification accuracy as a function of the quantization bit level. It can be observed that the reduction in quantization bit level leads to a severe degradation in classification performance. Moreover, despite the fact that lower bit levels allow for more effective data compression and significantly reduce communication requirements, the communication overhead still exceeds the efficiency achieved by the proposed AirComp method.

## 5 Conclusion

In this work, we proposed a multi-view over-the-air aggregation framework that integrates split learning and AirComp to address the communication bottleneck in multi-view sensing wireless networks. By splitting the ResNet18 model into sensor-side feature extraction and server-side classification, our method significantly reduces communication overhead while maintaining high classification accuracy. Experimental results on the ModelNet40 dataset demonstrate the effectiveness and robustness of our approach under noisy conditions.

Future work will focus on extending the framework to more complex datasets, exploring advanced denoising techniques, and further optimizing the split learning strategy to improve real-time performance in dynamic multi-view systems.

## Acknowledgments

The author would like to express their gratitude to Dr. David Lindell for his invaluable guidance and support throughout CSC2529: Computational Imaging. Furthermore, the author sincerely appreciate the valuable ideas and suggestions provided by Teaching Assistant Parsa Mirdehghan during the inception of this project.

## References

[1] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015.

[2] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, page 92–100, New York, NY, USA, 1998. Association for Computing Machinery.

[3] Chang Xu, Dacheng Tao, and Chao Xu. A Survey on Multi-view Learning, April 2013. arXiv:1304.5634 [cs].

[4] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Multi-view perceptron: a deep model for learning face identity and view representations. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, page 217–225, Cambridge, MA, USA, 2014. MIT Press.

[5] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page III–1247–III–1255. JMLR.org, 2013.

[6] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 689–696, Madison, WI, USA, 2011. Omnipress.

[7] Xu Chen, Khaled B. Letaief, and Kaibin Huang. On the View-and-Channel Aggregation Gain in Integrated Sensing and Edge AI. *IEEE Journal on Selected Areas in Communications*, 42(9):2292–2305, September 2024.

[8] Jiawei Shao, Yuyi Mao, and Jun Zhang. Task-oriented communication for multidevice cooperative edge inference. *IEEE Transactions on Wireless Communications*, 22(1):73–87, 2023.

[9] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.