Multi-View Latent Diffusion

Multi-View Latent Diffusion

G. Di Giacomo[1], G. Franzese[2], T. Cerquitelli[1], C. F. Chiasserini[1,3,4], P. Michiardi[2]
1: Politecnico di Torino, Italy – 2: EURECOM, France – 3: CNR-IEIIT, Italy – 4: CNIT, Italy

***Abstract*—Multi-view observations potentially offer a more comprehensive understanding of real-world phenomena compared to observations acquired from a single viewpoint. Existing models that utilize multi-view data often consider that all views are available during inference, but this assumption may not hold in practical scenarios. To address this limitation, we introduce MVLD, a novel method that, by employing a deterministic autoencoder and a score-based diffusion model, is capable of imputing missing views. We finally envision MVLD being used in a communication system for image transmission.**

I. Introduction

Real-world observations may be acquired from different viewpoints, providing a more comprehensive perception compared to single-view observations, which may not be sufficient to capture the complexity and diversity of a scene. Therefore, aggregating information from multiple views describing a given scene into a unified representation that can exploit *unique and redundant information* is a key objective.

Driven by these motivations, Multi-View Representation Learning is a growing research field that aims to find a meaningful representation from multi-view observations by learning both the correlation across views and their specific information [1]. Recently, multi-view datasets, e.g., [2]–[4], and Machine Learning models that rely upon them have received significant attention. Many of the available models exploit multi-view data to perform a specific task, such as 3D object reconstruction [5] or human pose estimation [6]. However, an important underlying assumption in the literature is that all input views are available at inference time. In practical deployments, such an assumption might fall short: sensor failure, obstructions due to dirt, and other phenomena are typical cases that must be addressed. To tackle this issue, some methods have been designed specifically to handle missing views [7], [8], or propose a preliminary phase to *conditionally* generate the missing views before performing the actual downstream task [9].

In this context, **our main contributions are twofold:** (i) we present and study a new method called Multi-View Latent Diffusion (MVLD) that enables conditional generation of any missing view at *inference time*, and (ii) we envision our method being used in a communication system for multi-view images transmission. Specifically, MVLD endows end-to-end systems with the ability to carry out data imputation, prior to performing a downstream task. Given a set of observed views, MVLD uses a deterministic autoencoder and a score-based diffusion model that operates on latent representations to generate the missing views. We study the quality of such generated views by computing the Fréchet Inception Distance (FID), a relevant metric used to assess the quality of synthetic images.

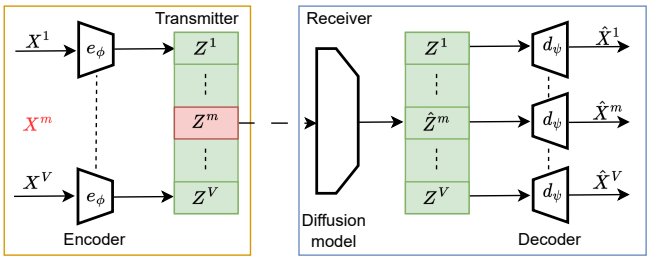

Fig. 1. Architecture of the proposed MVLD-based communication system for multi-view images transmission.

We investigate the use of MVLD in a scenario where such multi-view data must be transmitted over a wireless communication channel. If available views can be properly reconstructed from their latent representations and the conditional generation of the missing views is effective, one can think of proactively transmitting only the latent representation of a subset of views, from which it is possible to generate the missing ones at the receiver, for example, to save channel capacity. Therefore, we study the trade-off between the network load and the quality of generated images. Importantly, the trade-off depends on the characteristics of the transmission channel. For instance, if the network has low bandwidth, it may be convenient to send the latent variables of the least possible number of views that allow reconstructing the missing ones with sufficient quality.

Interestingly, MVLD can be employed for semantic communication. According to this paradigm, instead of transmitting the exact sequence of bits, only the semantic information, i.e., the meaning, of the data is transmitted and then used at the receiver to generate data *semantically equivalent* to the original ones. To assess the semantic preservation, one could evaluate the coherence of conditionally generated views; for instance, a pre-trained classifier can be employed both on imputed and observed views: if the class predictions are consistent the coherence is maintained.

II. Methodology

We build on the method described in [10], which deals with the general problem of modeling multiple input modalities that describe the same concept using, for example, image, audio, and text data. Specifically, missing views can be generated by following the same two-stage procedure presented in [10].

979-8-3503-2445-7/23/$31.00 ©2023 IEEE

Our approach is depicted in Fig. 1, where we consider V views in total and, for example, one missing view X^m. At the transmitter, a deterministic encoder e_ϕ is used to encode each observed view X^v, with $v{=}\{1,\dots,V\}\backslash m$, to obtain their latent representations $Z^v{=}e_\phi(X^v)$. Such latent representations are concatenated and sent to the receiver, where they are input to a score-based diffusion model, which we design for conditional generation through an original method that enables latent variables to evolve according to different arrows of time in the forward process, and that induces a correlation between latent variables in the backward process by means of a joint score network. The diffusion model can thus generate the latent variable $\hat{Z}^m$ of the missing view X^m, and, finally, a deterministic decoder d_ψ transforms available and generated latent variables back into the input space, thus obtaining $\hat{X}^v{=}d_\psi(Z^v), v{=}\{1,\dots,V\}\backslash m$, and $\hat{X}^m{=}d_\psi(\hat{Z}^m)$.

III. Background

In this section, we provide a brief overview of the two main components of MVLD, namely the deterministic autoencoder and the score-based diffusion model.

Autoencoder. The deterministic autoencoder is composed of two blocks, the encoder e_ϕ and the decoder d_ψ, and is trained separately and before the diffusion model. Denoting with $p(x)$ and l, respectively, the data distribution and a distance function, we train the autoencoder by minimizing the following loss:

$$\mathcal{L} = \int p(x) l(x - d_\psi(e_\phi(x)))\, dx\,. \tag{1}$$

Importantly, we use a deterministic autoencoder as it can guarantee no loss of information when mapping data into the latent space.

Score-based diffusion model. Once the deterministic autoencoder is trained, the encoder is employed to obtain the data latent representations, which are used for the training of the score-based diffusion model. During this stage, the diffusion model learns the distribution of such latent representations, which allows performing conditional generation during the inference phase.

In general, score-based generative modeling involves two steps, namely, the forward and the backward diffusion process. The former is a stochastic noising process, which injects noise into the input data, i.e., the latent representations in this case, while the latter reverses the noise perturbation. The forward process is defined by the following Stochastic Differential Equation (SDE)[1]:

$$\mathrm{d}R_t = \alpha(t) R_t \mathrm{d}t + g(t)\mathrm{d}W_t, \quad R_0 \sim q(r,0), \tag{2}$$

where R_t is the diffused random variable, while $\alpha(t)R_t$ and $g(t)$ denote the drift and the diffusion terms, respectively. W_t is a Wiener process and $q(r,t)$ denotes the probability density of the stochastic process at time $t{\in}[0,T]$; therefore, $R_0{\sim}q(r,0)$ is the initial condition influencing the noising process, where, in our case, $q(r,0)$ is the latent distribution.

To generate a new sample, we need to reverse the noising process; we thus derive the following reverse-time SDE[1]:

$$\begin{aligned}\mathrm{d}R_t = &\left(-\alpha(T{-}t)R_t + g^2(T{-}t)\nabla\log(q(R_t, T{-}t))\right)\mathrm{d}t \\ &+ g(T{-}t)\mathrm{d}W_t, \quad R_0 \sim q(r,T),\end{aligned} \tag{3}$$

which can be simulated by using a numerical integration scheme. We remark that, to do so, we first need to estimate the term $\nabla\log(q(R_t, T{-}t))$, i.e., the true score function, by using a parametric score network.

Since in this work we are interested in the conditional generation of the missing views, we properly modify (2) and (3) so that only the latent variables of the missing views are diffused. Also, we modify the true score function to make it conditioned on the observed views.

IV. Related work

As previously mentioned, our work relies on the approach presented in [10], which is designed to operate within the multi-modal domain. The study in [10] and, consequently, ours are related to the branch of works that employ combinations of Variational Autoencoders (VAEs) for generative modeling of multi-modal data, for both joint and conditional generation [11]–[13]. However, [10] demonstrates that these methods suffer from a trade-off between generative quality and coherence among modalities, i.e., it is not possible to improve one aspect without negatively affecting the other. Also, they are generally outperformed by the method in [10], which is the reason why we use such a method in our study.

The work in [14] identifies as the "generative learning trilemma" the three main requirements in generative modeling: high-quality sample generation, sample diversity, and fast sampling. This work underlines that methods based on Generative Adversarial Networks (GANs) suffer from poor mode coverage, that is the diversity of generation is bad, and that VAEs suffer from poor image quality. Score-based diffusion models, instead, generate both high-quality and diverse images, but they are slower at sampling. To tackle this issue, they present a denoising diffusion GAN, whose performance in terms of sample quality and diversity is comparable to standard diffusion models, while achieving a much faster sampling. The work in [15] provides a rigorous analysis of diffusion times in score-based generative models and presents a new method that, by adopting smaller diffusion time values, is more computationally efficient in both training and sampling, while achieving competitive or higher performance compared to standard diffusion models and other competitors. To accelerate sampling, the work in [16] introduces a faster numerical integration scheme.

In the literature, a growing number of papers investigate multi-view clustering, a method relying on common and specific information from multiple views to partition data into clusters. Some works specifically focus on partial multi-view clustering, where data are not assumed to be complete; for instance, [17], [18] address the problem of missing views

[1]We use the same notation as the one adopted in [10].

by imputing them using a GAN. Similarly, [9] generates the missing views before performing a classification task.

Our paper is related also to another research area, namely, generative modeling for communications and particularly for semantic communications. [19] proposes a GAN-based semantic communication system for image transmission that allows drastic data compression while achieving high-quality image generation. More precisely, the transmitter employs an encoder to compute the images latent representations, which are sent to the receiver where they are fed to a GAN generator to restore the image content. A second method also uses a heatmap and a semantic (or instance) map, obtained with a pre-trained segmentation model, to specify the regions that can be completely generated and those whose content must be preserved. An approach similar to the latter is adopted in [20], while [21] employs a diffusion model to generate images starting from one-hot encoded maps; such maps are obtained at the sender by applying a segmentation model and then they are transmitted over the communication channel. [22] leverages the generator of a GAN for image semantic communication. First, the sender maps the images to their latent representations using the GAN inversion method; then, the latent variables are transmitted to the receiver, which reconstructs the image using the generator.

V. Conclusions

In this paper, we discussed the relevance and the benefits of multi-view data, and the fact that, in practical scenarios, such multi-view data may present missing views at inference time. To address this drawback, we introduced the MVLD method to generate missing views conditioned on the observed ones. We then argued that MVLD can be used in a communication system for multi-view images transmission, allowing the imputation of potential missing views. For future work, we will perform extensive experiments on multiple datasets to test MVLD performance for conditional generation and to evaluate its robustness when deployed in a communication system. Finally, we will extend our work with a faster numerical integration scheme, and study the applicability of recent variant of diffusion models [23] that operates in the function space, which would allow our model to be scale- and resolution-free, while using simpler architectures to learn the score network.

Acknowledgment

This work was supported by the SNS-JU-2022 project ADROIT6G under the European Union's Horizon Europe research and innovation programme under Grand Agreement No. 101095363.

References

[1] H. Hwang, G.-H. Kim, S. Hong *et al.*, "Multi-view representation learning via total correlation objective," in *NeurIPS*, 2021.
[2] H. Joo, H. Liu, L. Tan *et al.*, "Panoptic studio: A massively multiview system for social motion capture," in *ICCV*, 2015.
[3] Z. Yu, J. Shin Yoon, I. K. Lee *et al.*, "Humbi: A large multiview dataset of human body expressions," in *CVPR*, 2020.
[4] T. Chavdarova, P. Baqué, S. Bouquet *et al.*, "Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection," in *CVPR*, 2018.
[5] D. Wang, X. Cui, X. Chen *et al.*, "Multi-view 3d reconstruction with transformers," in *ICCV*, 2021.
[6] J. Zhang, Y. Cai, S. Yan *et al.*, "Direct multi-view multi-person 3d pose estimation," in *NeurIPS*, 2021.
[7] Q. Tan, G. Yu, C. Domeniconi *et al.*, "Incomplete multi-view weak-label learning." in *IJCAI*, 2018.
[8] A. Kanezaki, Y. Matsushita, and Y. Nishida, "Rotationnet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints," in *CVPR*, 2018.
[9] M. Xie, Z. Han, C. Zhang *et al.*, "Exploring and exploiting uncertainty for incomplete multi-view classification," in *CVPR*, 2023.
[10] M. Bounoua, G. Franzese, and P. Michiardi, "Multi-modal latent diffusion," *arXiv preprint arXiv:2306.04445*, 2023.
[11] M. Wu and N. Goodman, "Multimodal generative models for scalable weakly-supervised learning," in *NeurIPS*, 2018.
[12] Y. Shi, S. N, B. Paige *et al.*, "Variational mixture-of-experts autoencoders for multi-modal deep generative models," in *NeurIPS*, 2019.
[13] T. M. Sutter, I. Daunhawer, and J. E. Vogt, "Generalized multimodal ELBO," in *ICLR*, 2021.
[14] Z. Xiao, K. Kreis, and A. Vahdat, "Tackling the generative learning trilemma with denoising diffusion GANs," in *ICLR*, 2022.
[15] G. Franzese, S. Rossi, L. Yang *et al.*, "How much is enough? a study on diffusion times in score-based generative models," *Entropy*, vol. 25, no. 4, p. 633, 2023.
[16] A. Jolicoeur-Martineau, K. Li, R. Piché-Taillefer, T. Kachman *et al.*, "Gotta go fast when generating data with score-based models," *arXiv preprint arXiv:2105.14080*, 2021.
[17] C. Zhang, Y. Cui, Z. Han *et al.*, "Deep partial multi-view learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
[18] C. Xu, Z. Guan, W. Zhao *et al.*, "Adversarial incomplete multi-view clustering." in *IJCAI*, 2019.
[19] E. Agustsson, M. Tschannen, F. Mentzer *et al.*, "Generative adversarial networks for extreme learned image compression," in *ICCV*, 2019.
[20] D. Huang, X. Tao, F. Gao *et al.*, "Deep learning-based image semantic coding for semantic communications," in *GLOBECOM*, 2021.
[21] E. Grassucci, S. Barbarossa, and D. Comminiello, "Generative semantic communication: Diffusion models beyond bit recovery," *arXiv preprint arXiv:2306.04321*, 2023.
[22] T. Han, J. Tang, Q. Yang *et al.*, "Generative model based highly efficient semantic communication approach for image transmission," in *ICASSP*, 2023.
[23] G. Franzese, G. Corallo, S. Rossi *et al.*, "Continuous-time functional diffusion processes," *arXiv preprint arXiv:2303.00800*, 2023.