TRANSLATION-BASED MULTIMODAL LEARNING

by

ZHENGYI LU

A thesis submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN ELECTRICAL AND COMPUTER ENGINEERING

2024

Oakland University
Rochester, Michigan

Master Advisory Committee:

Jia Li, Ph.D., Chair
Shadi Alawneh, Ph.D.
Wing-Yue Geoffrey Louie, Ph.D.

To my family

ACKNOWLEDGMENTS

The story of my journey at Oakland University began in the fall of 2023, and this thesis reaches its completion in the fall of 2024. I am excited to continue my PhD studies at Oakland University, and I hope to make even more wonderful memories in the years to come.

I would like to express my heartfelt gratitude to all those who have helped me along the way. First and foremost, I want to thank my supervisor, Dr. Jia Li, who truly opened the door to research for me. Her patient guidance, encouragement, and extensive knowledge have been invaluable, filling me with enthusiasm and confidence for scientific exploration. I would also like to thank Dr. Lianxiang Yang and Dr. Gary Barber for their dedication to the exchange program, which gave me a life-changing opportunity. Similarly, I extend my gratitude to Dr. Wing-Yue Geoffrey Louie and Dr. Shadi Alawneh, members of my thesis committee, for their support in my research. I would also like to thank Dr. Steven Louis for providing the LaTeX template that made this thesis possible. I also want to thank all the friends who have accompanied me along this journey, for their companionship and support through both the challenges and the joyful moments.

Finally, my deepest appreciation goes to my family. Your unwavering support has given me the strength to stand on your shoulders and see the world. Thank you for always being there for me without complaint, for your silent support behind the scenes, and for respecting my decisions. I couldn't have done it without you.

See you all when the maple leaves turn red four more times.

ZHENGYI LU

ABSTRACT

TRANSLATION-BASED MULTIMODAL LEARNING

by

ZHENGYI LU

Adviser: Jia Li, Ph.D.

Multimodal learning has become a critical area of research in artificial intelligence, aiming to effectively integrate and translate information across different data modalities such as images, text, and audio. However, existing approaches often struggle with data scarcity and robustness when modalities are incomplete or missing. To address this gap, this work investigates translation-based multimodal learning through two complementary approaches: xDSBMIT and TransTrans, corresponding to end-to-end and representation-level translation methods.

The xDSBMIT framework integrates the Diffusion Schrödinger Bridge (DSB) with the diffusion process, offering an effective solution for multimodal image translation, specifically applied to Synthetic Aperture Radar (SAR) to Electro-Optical (EO) and Infrared (IR) image translation. TransTrans addresses multimodal sentiment analysis through representation-level translation, reconstructing missing modalities in real-time using a Transformer-based architecture. In experiments, xDSBMIT achieved high-quality translations in SAR2IR and SAR2EO tasks with limited datasets, significantly outperforming traditional methods. TransTrans demonstrated superior performance in sentiment analysis under missing modality conditions. Overall, xDSBMIT and TransTrans provide complementary solutions to challenges in translation-based multimodal learning, advancing the state-of-the-art in image translation and sentiment analysis.

TABLE OF CONTENTS

TABLE OF CONTENTS—Continued

TABLE OF CONTENTS—Continued

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI Artificial Intelligence

DSB Diffusion Schrödinger Bridge

EO Electro-Optical

GAN Generative Adversarial Network

IR Infrared

NLP Natural Language Processing

SAR Synthetic Aperture Radar

xDSBMIT Explainable Diffusion Schrödinger Bridge Model for Image Translation

CHAPTER ONE

INTRODUCTION

1.1 Translation-based Learning

With the rapid development of deep learning, multimodal Learning has become a research hotspot in the field of Artificial Intelligence (AI). Among these, translation, as a core technology, has achieved great success in Natural Language Processing (NLP) [14–17]. Traditional text translation aims to achieve information conversion between different languages, which is essentially data mapping between different domains. The extension of this concept allows us to perform image-to-image translation between different visual domains, such as style transfer and cross-sensor data conversion.

Initially, Neural Machine Translation achieved high-quality translation from source language to target language through encoder-decoder architectures [15] and attention mechanisms [18]. This success inspired researchers to apply similar architectures to the image domain, proposing encoder-decoder-based Image-to-Image Translation [19]. Some researchers use the Generative Adversarial Network (GAN) [20] framework to achieve this, such as using cGAN [21] to assist translation with additional information or using CycleGAN [22] for unpaired translation. Others employ the VAE [23] framework to learn the probability distribution of data for high-quality image translation. The goal of both methods is to learn the mapping from one image domain to another while preserving the fundamental content of the images. For example, translating label maps into photographs [13], or converting daytime cityscapes into nighttime scenes [24].

Furthermore, the concept of image translation has been extended to image conversion between different styles and sensors. Style Transfer allows us to combine the content of one image with the style of another, creating unique artistic effects [25, 26]. In

fields such as remote sensing and medical imaging, different sensors may capture different features of the same scene. Through translation models, we can achieve mutual generation between these different images. [27–31].

Meanwhile, Translation-Based Learning has been introduced into conversions between other modalities, such as Image-to-Text [32–35], Audio-to-Text [36, 37], Audio-to-Image [38], Image-to-Audio [39], and Text-to-Image [40]. The generated results from these modality translations are new modalities, and the current needs of multimodal learning have further developed, requiring repeated modality fusion after generating new modalities.

1.2 Multimodal Learning

As the limitations of translation learning between modalities became increasingly evident, researchers have sought to enhance these models by combining multimodal learning [41] with the rapidly evolving field of representation learning [42]. This integration allows for a deeper understanding of each modality's unique features—be it text, visual, or audio data—by representing them at the feature level before fusing the learned representations. By utilizing shared representations, models can bridge the gap between different data modalities, leading to more effective and complementary information exchange across domains. This approach has enabled significant improvements in tasks that require the integration of multiple modalities, such as machine translation, image-to-text conversion, and audio-visual sentiment analysis.

Multimodal translation-based learning capitalizes on this concept by not only enabling the mapping of information across modalities but also allowing the underlying models to learn the intrinsic correlations and dependencies between these modalities. Unlike conventional unimodal models, which focus on translating one type of data (e.g., text-to-text or image-to-image), multimodal translation-based learning aims to translate

information between fundamentally different types of data. For example, the translation of text into corresponding images or the generation of audio descriptions from video inputs. By pre-representing the text, visual, and audio modalities and then merging these features at a shared representation level, the resulting models can more effectively harness the strengths of each modality, thereby improving the overall task performance.

One of the most significant applications of this approach is Multimodal Sentiment Analysis [43–46]. In sentiment analysis, where emotion or opinion must be extracted from data, relying on a single modality can be insufficient. For example, speech signals may be distorted by excessive background noise, video frames might be occluded by visual distractions, and textual data may be ambiguous or lack contextual depth. In such scenarios, translation-based learning across multiple modalities becomes crucial. Through multimodal fusion, these models can combine the information provided by complementary data types—such as using video cues to supplement missing information from audio or text. This process leads to a more robust analysis, enabling the system to mitigate issues like noise, missing data, or ambiguities from individual modalities, thereby improving the accuracy and robustness of sentiment recognition [47–53].

Multimodal translation-based learning, therefore, represents a significant leap forward in the broader field of artificial intelligence, particularly for applications that require rich, contextual understanding of diverse datasets. Whether it's transforming images into detailed textual descriptions, generating audio from visual inputs, or enhancing sentiment analysis through multimodal fusion, this approach opens new avenues for achieving higher performance in tasks that benefit from cross-modal learning. By bridging the gaps between different types of sensory inputs, translation-based learning not only addresses the limitations of unimodal systems but also unlocks the potential to better capture the complex relationships inherent in multimodal data.

1.3 Key Contribution

The two key approaches explored in this thesis, Explainable Diffusion Schrödinger Bridge Model for Image Translation (xDSBMIT) and TransTrans, represent distinct advancements in the field of translation-based multimodal learning, each catering to unique challenges within specific application domains.

The first key attempt at end-to-end translation explored in this thesis is the Diffusion Schrödinger Bridge (DSB) framework. DSB integrates the diffusion process with the Schrödinger Bridge problem, a method that enhances both stability and interpretability in multimodal image translation. The diffusion model progressively transforms one image domain into another by leveraging the unique characteristics of the image distributions. Applied to Synthetic Aperture Radar (SAR) to Electro-Optical (EO) and Infrared (IR) image translation, this approach enables efficient, high-quality translations even with limited datasets, offering significant improvements in remote sensing applications.

The second key approach, focusing on representation-level translation, is the TransTrans framework. Unlike xDSBMIT, which focuses on an end-to-end translation mechanism, TransTrans addresses multimodal sentiment analysis by leveraging translation-based learning to handle missing modalities in real-time. This Transformer-based system incorporates a translation-driven mechanism that reconstructs missing modalities, such as predicting missing visual data from audio and text features. This not only enhances the robustness of multimodal learning but also improves sentiment prediction accuracy, making it a powerful framework for scenarios with incomplete data.

The xDSBMIT framework excels in terms of interpretability and performs remarkably well with limited data. By integrating the Diffusion Schrödinger Bridge with multimodal image translation, xDSBMIT leverages the unique characteristics of different modalities to achieve high-quality translations, even with small datasets. On the other

hand, the TransTrans framework combines Transformer architecture with various pre-trained models, tailoring its feature extraction methods to the specific characteristics of each modality. This approach ensures that each modality—whether it be audio, text, or visual data—is processed optimally, resulting in more robust and accurate sentiment analysis, particularly in scenarios where data may be incomplete.

These two models demonstrate how translation-based learning can be applied across diverse fields, from image translation in remote sensing to sentiment analysis in social media and customer feedback. By addressing both data scarcity and multimodal robustness, xDSBMIT and TransTrans provide complementary solutions that push the boundaries of current multimodal learning research.

- xDSBMIT (Multimodal Image Translation):
  - High interpretability and stability in image translation tasks.
  - Effective in translating between SAR, EO, and IR images.
  - Performs exceptionally well with limited datasets, reducing the need for large-scale labeled data.
  - Suitable for remote sensing and satellite imagery applications.
  - Utilizes the Diffusion Schrödinger Bridge to achieve stable transformations across image modalities.
- TransTrans (Multimodal Sentiment Analysis):
  - Integrates Transformer architectures with pre-trained models for modality-specific feature extraction.
  - Specializes in handling audio, text, and visual data for sentiment analysis tasks.
  - Improves robustness by reconstructing missing modalities in real-time.

- Enhances sentiment prediction accuracy, especially in scenarios with incomplete or noisy data.
- Focuses on the unique characteristics of each modality, ensuring optimal feature processing.

Together, these two frameworks reflect the power of translation-based multimodal learning, offering innovative solutions across different domains. As we move forward, the next sections will explore the detailed architecture and experimental evaluation of these models, providing insight into their performance and contributions to the broader field of multimodal learning.

CHAPTER TWO

LITERATURE REVIEW

2.1 END-TO-END TRANSLATION

End-to-end translation methods aim to directly map input from one modality to another in a fully integrated system, where the entire translation process is trained jointly without intermediate steps. These methods leverage deep learning techniques to learn complex transformations between modalities such as image-to-image, text-to-image, and audio-to-text translation. They have become increasingly significant in a variety of applications, ranging from artistic content generation to practical uses in medical imaging and autonomous systems.

One of the most successful approaches in end-to-end translation is the encoder-decoder architecture, which was initially popularized in neural machine translation (NMT) tasks [16]. In this framework, the encoder transforms the source modality into an intermediate representation, which is then passed through a decoder to generate the target modality. These architectures have been extended beyond text-based tasks to applications such as image-to-image translation [13], audio-to-text translation [36, 37], and other cross-modal tasks. An important enhancement in this architecture is the incorporation of attention mechanisms [18], which dynamically focus on relevant parts of the input during translation, significantly improving performance in both natural language and image domains. The attention mechanism has also facilitated the handling of long-range dependencies, enabling more accurate translation even with complex and high-dimensional input data.

Another significant end-to-end translation approach is the use of conditional generative adversarial networks (cGANs), which incorporate additional information, such

Figure 2.1: Pix2Pix Framework. Source: [13]

as labels or other conditional data, to guide the generation process [21]. Unlike standard GANs, cGANs have proven effective in generating high-quality outputs in tasks like text-to-image and image-to-image translation. One of the pioneering models in image-to-image translation is Pix2Pix, as shown in Fig 2.1, which is built on a conditional GAN framework and is specifically designed for tasks with paired data [13]. Pix2Pix learns the mapping from one domain to another by leveraging the known correspondence between input and target images, and has been used for translating sketches into realistic images, colorizing grayscale images, and converting semantic label maps into photorealistic scenes. Pix2PixHD extends this approach by enabling high-resolution image generation, making it suitable for more detailed and photorealistic outputs [24]. These models have been instrumental in advancing applications in content creation, artistic style transfer, and realistic image synthesis. When paired data is not available, CycleGAN introduces an innovative solution by enforcing cycle consistency [22]. CycleGAN allows training in scenarios where paired training data is unavailable, making it applicable to a wide range of tasks such as style transfer, cross-sensor image translation, and medical imaging [27]. This model addresses the challenge of unpaired translation by ensuring that an image translated to another domain and back yields the original image, thus maintaining consistency across translations. The cycle consistency constraint has

made CycleGAN a powerful tool for many real-world applications where paired data is challenging to collect, such as translating between artistic styles or enhancing satellite images.

Variational Autoencoders (VAEs) provide yet another approach to end-to-end translation by learning the latent probability distributions of the input modalities [23]. VAEs are effective in generating outputs that resemble the input data distribution, making them suitable for tasks like image generation and style transfer. By learning a latent space that can be sampled to generate new data, VAEs have found applications in synthetic medical image generation and translating between different imaging modalities. Their probabilistic nature allows VAEs to generate diverse outputs, making them useful for scenarios where output variability is desired, such as in data augmentation for training other deep learning models.

The Recurrent Multistage Fusion Network (RMFN) uses recurrent neural networks (RNNs) to perform multistage fusion across different modalities, refining features at each stage to capture both local and global dependencies [54]. This approach has been particularly successful in tasks requiring temporal context, such as video-to-text or speech-to-text translation, demonstrating the effectiveness of end-to-end models in handling sequential data. By refining features across multiple stages, RMFN captures intricate temporal patterns, making it highly effective for applications in video summarization, human activity recognition, and multimedia analysis.

Transformer-based models have recently gained prominence in multimodal end-to-end translation tasks, where their self-attention mechanisms enable capturing long-range dependencies between different modalities [17]. Transformers are highly effective in tasks such as video-to-text translation and audio-to-image generation, achieving state-of-the-art results due to their parallel processing capabilities and attention layers. Unified model architectures, such as GPT-3 and BERT-based multimodal

transformers, aim to integrate multiple tasks and modalities into a single framework, simplifying the training process and improving performance across various tasks. These models utilize shared embeddings and attention layers to process multiple modalities in an end-to-end fashion, paving the way for highly generalizable models that can handle a wide range of translation tasks [55]. The flexibility and scalability of transformers have made them an indispensable tool in tasks involving large datasets and complex multimodal interactions, such as in virtual assistants and autonomous systems.

Diffusion-based models have also gained attention for their effectiveness in modality translation. In paired translation tasks, diffusion models iteratively denoise latent representations to match the target modality, ensuring smooth transitions between modalities. Brownian Bridge-based diffusion methods have been particularly effective in image coloring, where the model refines an initial noisy grayscale image into a fully colored output through a series of stochastic transformations [56]. More advanced approaches, such as the Diffusion Schrödinger Bridge (DSB) model, have been applied to image-to-image (I2I) translation, learning to map between source and target modalities by solving a Schrödinger Bridge problem, which improves image quality and provides interpretable dynamics of the diffusion process [57, 58]. Additionally, recent research from Tsinghua University has extended the Schrödinger Bridge framework to speech translation tasks, enabling robust cross-modal translations between speech and text by leveraging the smoothness and flexibility of diffusion processes, achieving state-of-the-art results in multilingual speech translation [59]. The ability of diffusion-based models to provide smooth and consistent transformations between modalities makes them a promising direction for future research in tasks involving intricate cross-modal dynamics, such as multi-language speech synthesis and detailed medical imaging translations.

An important recent development in end-to-end translation is DALL-E, an advanced text-to-image model that represents a significant breakthrough in generating

photorealistic images based on textual descriptions [40]. By using a transformer-based architecture, DALL-E is capable of capturing intricate relationships between objects and scenes, highlighting the potential of multimodal learning in creative tasks where input modalities are vastly different. The success of DALL-E showcases the power of deep learning in bridging the gap between text and visual understanding, opening new possibilities in areas such as graphic design, content creation, and even assisting in complex design processes across industries.

2.2 Representation-level Translation and Its Applications

Representation-level translation aims to learn shared latent spaces between different modalities, thereby enabling models to effectively fuse and align information at a representation level. This approach offers enhanced robustness to noise and missing data, providing flexibility for a wide range of applications, including multimodal sentiment analysis, medical imaging, cross-modal retrieval, and text-to-image generation. Compared to end-to-end translation methods, representation-level approaches offer superior adaptability and generalizability.

Effective representation-level translation requires precise alignment of the latent spaces of different modalities. He et al. [60] demonstrated that masked autoencoders can learn scalable latent space representations, facilitating the alignment between available and missing modalities. However, while masked autoencoders offer significant improvements, they may struggle with capturing complex inter-modal relationships. To address this, graph-based fusion approaches have emerged as an effective solution that can capture complex dependencies between modalities through graph-based structures.

Graph-based fusion has emerged as another significant approach for multimodal fusion, effectively modeling relationships between different modalities by representing them as nodes in a graph. Bischke et al. [61] illustrated the efficacy of this method in

building segmentation tasks. By representing each modality as a node, the message-passing algorithm integrates the available data into the latent representation space, ensuring that the model can effectively handle missing modalities. Despite its effectiveness, graph-based approaches are often computationally intensive, which can be a limitation in large-scale applications. This has led to growing interest in using more computationally efficient methods, such as Variational Autoencoders (VAEs), for handling missing modalities.

Variational Autoencoders (VAEs) have also been employed to generate representations for missing modalities by learning the latent distribution of the available data. Hamghalam et al. [62] demonstrated that VAEs can impute missing information in medical segmentation tasks by leveraging learned latent representations, thereby generating accurate segmentation results. While VAEs provide a probabilistic framework for handling missing data, they may still struggle with modeling highly structured data or capturing long-range dependencies between modalities. To address this, hierarchical encoder-decoder structures have been proposed as an extension to VAEs to better capture complex data distributions.

Li et al. [63] proposed the use of hierarchical encoder-decoder structures to generate missing modality representations, enabling downstream tasks to benefit from a more comprehensive multimodal representation space. Although this hierarchical approach enhances the ability to model complex data, it often requires extensive computational resources and careful tuning. This limitation has motivated researchers to explore more efficient architectures that can still effectively model inter-modal relationships, such as the Transformer.

The Transformer architecture has also been effectively adapted to representation-level translation. Tsai et al. [55] proposed a multimodal transformer model that aligns features across modalities by mapping them into a common latent space. This

approach allows for the effective processing of unaligned multimodal language sequences, demonstrating the utility of representation-level translation in handling missing data in natural language processing tasks. However, Transformer models can be sensitive to modality alignment errors and require a large amount of training data. These limitations have driven the development of alternative approaches like cross-modal fusion, which directly leverages shared latent spaces to improve alignment robustness.

In medical imaging, Zhou et al. [64] and Sun et al. [65] explored cross-modal fusion for brain tumor segmentation. By leveraging a shared latent space across MRI modalities, their models effectively address the issue of missing modalities, leading to improved segmentation performance. However, achieving well-aligned latent spaces remains challenging, particularly when dealing with diverse and heterogeneous datasets. To overcome this challenge, relation-aware approaches have been introduced to explicitly learn correlations between available modalities.

For tasks such as Audio-Visual Question Answering (AVQA), Park et al. [66] proposed a relation-aware missing modality generator that learns latent correlations between modalities to predict missing features. This relation-aware approach enhances the robustness of AVQA systems, making it a strong candidate for addressing modality incompleteness in multimedia tasks. However, relation-aware models can be computationally expensive and complex to train due to the need to learn intricate relationships across multiple modalities. To simplify the training process while retaining effectiveness, researchers have turned to self-supervised joint embeddings.

Kim et al. [67] introduced a self-supervised joint embedding architecture that employs predictive learning to generate missing modality features. This self-supervised approach aligns representations from existing modalities with those of missing modalities without requiring end-to-end supervision, thus making it efficient for tasks involving incomplete data. While self-supervised methods are often simpler to train, they may

sometimes produce suboptimal feature quality, particularly in highly complex multimodal scenarios. To address these quality issues, prompt learning techniques have been utilized to improve the generation of high-quality latent features.

Prompt learning has also emerged as an effective technique for missing modality generation. Guo et al. [68] developed a prompt learning framework that maps available modality prompts into the latent space to generate representations for missing modalities. By focusing solely on the representation level, this method reduces training complexity while maintaining high performance, making it well-suited for multimodal tasks with limited data availability. However, prompt learning approaches may not always generalize well to new or unseen tasks, leading to the development of more adaptable models, such as the U-Adapter, that can stabilize the integration of missing data.

Lin et al. [69] proposed the U-Adapter model for cross-modal fusion, which enables stable integration of missing modality data by preventing domain shifts in the latent space. The U-Adapter ensures that even when certain modalities are missing, the latent space remains well-aligned, thereby improving the performance of downstream tasks such as classification and segmentation. Although the U-Adapter provides a stable solution for integrating missing modalities, further research is needed to explore its adaptability across a broader range of domains and applications.

Collectively, these approaches demonstrate the versatility and effectiveness of representation-level translation across various applications, underscoring the benefits of robust latent space alignment, graph-based fusion, hierarchical learning, and prompt-based generation methods. By focusing on a shared latent space representation, these models achieve enhanced adaptability and resilience when dealing with incomplete multimodal data. Importantly, none of the original methods are removed or deleted but rather built upon to further advance the field, ensuring the retention of all foundational contributions. Moreover, the continuous development of novel techniques seeks to address

the limitations of previous methods, ensuring a more comprehensive and flexible framework for representation-level translation in complex multimodal environments.

CHAPTER THREE

DATASETS FOR TRANSLATION-BASED MULTIMODAL LEARNING

This section introduces several datasets commonly used in translation-based multimodal learning. These datasets span various tasks, including image-to-image translation, cross-modal sentiment analysis, and audio-to-text translation, providing comprehensive resources for training and evaluating multimodal learning models.

3.1 Cityscapes [1]

Cityscapes dataset is widely used for urban scene understanding, particularly in tasks such as semantic segmentation and image-to-image translation. It contains high-resolution street scene images from 50 cities, with fine-grained annotations of 30 object classes. This dataset is particularly useful for tasks involving street view style transfer, such as transforming daytime images into nighttime scenes, or translating between different weather conditions.

Figure 3.1: Cityscapes Dataset. Source: [1]

3.2 CMPFacades [2]

CMPFacades dataset consists of architectural facade images and their corresponding labels. It is used extensively in image-to-image translation tasks that involve architectural design, such as generating facade layouts or transforming the appearance of buildings. This dataset is also valuable for tasks such as architectural style transfer and facade completion, where models learn to generate realistic building facades from simple line drawings.

3.3 Aachen Day-Night [3]

Aachen Day-Night dataset includes urban images captured at different times of the day, making it ideal for day-to-night translation tasks. This dataset contains pairs of images captured during the day and at night, which are useful for research in cross-sensor data translation and enhancing night-time visual understanding in autonomous driving systems. The dataset's emphasis on varying lighting conditions enables robust model training for domain adaptation between different lighting environments.

3.4 UNICORN 2008 [4]

UNICORN 2008 dataset features multimodal data from Wide Area Motion Imagery (WAMI) and Synthetic Aperture Radar (SAR) sensors, as shown in Fig 3.2. It is specifically designed for tasks that require simultaneous alignment of visual and radar-based information. The dataset contains large format electro-optical (EO) sensor images and SAR frames, captured at approximately 2 frames per second. Due to the misalignment in time between EO and SAR frames, this dataset poses unique challenges for sensor fusion and cross-modal translation tasks, such as radar-to-image translation and vice versa.

Figure 3.2: UNICORN 2008. Source: [4]

3.5 Wikiart [5]

Wikiart dataset contains a vast collection of artwork images, organized by style, genre, and artist. It is widely used in style transfer tasks, where the goal is to apply artistic styles from famous paintings to real-world images. The dataset spans various artistic movements, providing models with the ability to learn style representations and apply them to different content images. Wikiart is crucial for research in creative image generation and cross-modal art synthesis.

3.6 Flickr30k [6]

Flickr30k dataset provides a large set of images paired with descriptive text annotations. This dataset is commonly used in vision-and-language tasks such as image

captioning, text-to-image generation, and cross-modal retrieval. With over 30,000 images and detailed text descriptions, models can learn the relationship between visual content and its textual representation, enabling the generation of textual descriptions from images and vice versa.

3.7 MS COCO [7]

MS COCO dataset is a large-scale dataset widely used in multimodal learning, particularly for tasks involving object detection, image segmentation, and image captioning. It includes over 330,000 images with rich annotations, making it a versatile dataset for both vision-only and vision-and-language tasks. In translation-based learning, MS COCO is frequently employed in text-to-image and image-to-text translation tasks.

3.8 Places Audio Caption [8]

Places Audio Caption dataset combines visual and audio data, allowing for tasks such as image-to-audio and audio-to-image translation. This dataset contains audio descriptions of various scenes, providing a unique resource for training models that translate between auditory and visual modalities. It is commonly used in research on multimodal fusion and cross-modal translation between sound and imagery.

3.9 AudioSet [9]

AudioSet is a large-scale dataset of labeled audio events, containing over 2 million human-labeled audio clips spanning more than 600 categories. This dataset is highly valuable for multimodal translation tasks, especially in audio-to-text and audio-to-visual translation. Models trained on AudioSet can learn to translate audio events into textual descriptions or generate corresponding visual scenes based on sound.

Figure 3.3: CMU-MOSI Dataset. Source: [10]

3.10 CMU-MOSI [10]

CMU-MOSI dataset is a multimodal sentiment analysis corpus that includes video, audio, and text modalities. As we can see in Fig 3.3, it consists of 2,199 opinion segments from YouTube videos, annotated for sentiment intensity on a continuous scale. CMU-MOSI is widely used in sentiment analysis tasks where models must fuse information from multiple modalities to predict sentiment polarity.

3.11 CMU-MOSEI [11]

CMU-MOSEI dataset extends CMU-MOSI with a larger collection of multimodal sentiment data. It includes over 23,000 opinion segments from 1,000 speakers, covering various topics. CMU-MOSEI provides sentiment and emotion annotations across text,

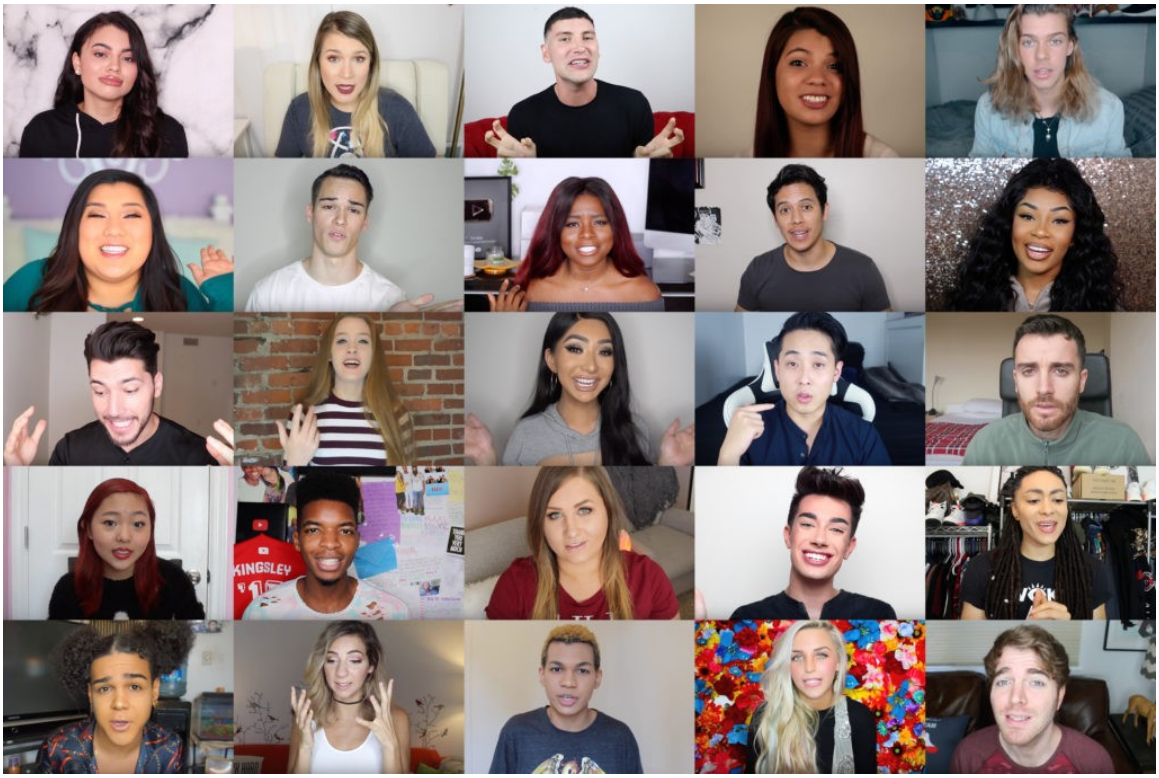

Figure 3.4: CMU-MOSEI Dataset. Source: [11]

audio, and video, making it ideal for tasks that involve multimodal emotion recognition and sentiment analysis.

3.12 IEMOCAP [12]

IEMOCAP dataset is a multimodal dataset created for emotion recognition tasks. It contains audiovisual recordings of actors performing improvised and scripted dialogues, with annotations for emotion categories such as anger, happiness, sadness, and neutral. IEMOCAP is frequently used for emotion recognition tasks that require the fusion of visual, auditory, and textual information.

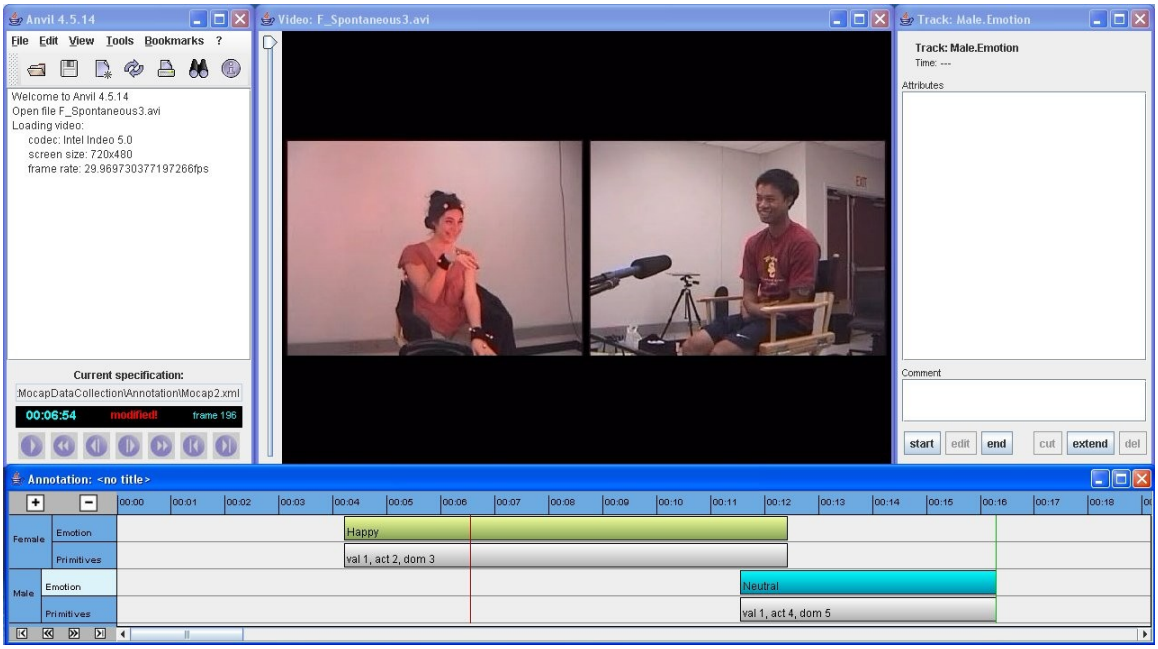

Figure 3.5: IEMOCAP. Source: [12]

CHAPTER FOUR

xDSBMIT:AN END-TO-END TRANSLATION

4.1 Image-to-Image Translation

Image-to-Image (I2I) translation involves converting images from one domain to another, leveraging techniques like style transfer, image colorization, super-resolution, and image synthesis [70–72]. Recently, this field has extended to multimodal learning, enabling translations across different modalities by training on extensive datasets [73]. These advancements have applications in artistic creation, medical imaging, and satellite image analysis, broadening the spectrum of image translation and enhancing the interpretation of visual data across various contexts.

Dynamic Data Driven Applications Systems (DDDAS) integrate instrumentation data with models in real-time, allowing these models to dynamically manage the use and acquisition of data. DDDAS-based methods adapt to the ever-changing nature of real-world systems, providing a robust and flexible framework for various applications that demand real-time data integration and dynamic system adaptability. Generative Adversarial Networks (GANs) are extensively used for image translation in dynamic data-driven applications systems (DDDAS) to generate augmented data for near-real support to deployed systems [74]. Notable techniques include pix2pix for paired image translation and CycleGAN for unpaired image translation [13, 75]. GAN-based methods use continuous feedback from the discriminator to produce realistic images closely approximating the ground truth, significantly advancing the field of image translation [76–78]. However, GANs face challenges like training difficulties, vanishing gradients, and poor interpretability.

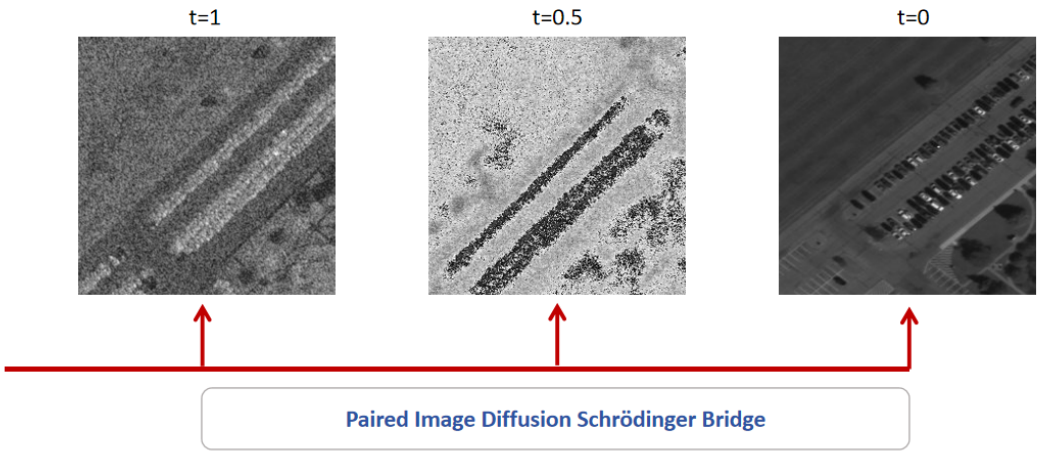

Figure 4.1: Image Generating via Diffusion Schrödinger bridge

To overcome these issues, diffusion-based methods for image translation have been explored [79]. Traditional diffusion models progressively denoise an image until it transforms into Gaussian noise, learning the reverse process to generate new images [80, 81]. However, these models struggle with paired image translation. To address this, the Diffusion Schrödinger Bridge (DSB) has been proposed as an innovative generative model, combining the Schrödinger Bridge framework with the diffusion process to transition smoothly between distributions and generate high-quality images.

DSB has shown promising results in various applications. De Bortoli explored its use in score-based generative modeling, demonstrating its versatility [82]. Liu demonstrated its effectiveness in I2I translation [**?**], and Chen highlighted its superiority in text-to-speech synthesis compared to traditional diffusion models [**?**]. However, previous applications of DSB have focused on image restoration and colorization, not on cross-modal image translation tasks. Additionally, there has been no practical implementation of DSB in the field of satellite imagery. To address the cross-modal image translation gap, and achieve stable, highly interpretable image translation, we propose the

application of the diffusion Schrödinger bridge in multimodal image translation tasks. As shown in Fig. 4.1, the red arrows in the figure represent the sampling process, the Paired Image Diffusion Schrödinger Bridge (DSB) can transform a noisy image at $t = 1$ into a clear image at $t = 0$ through intermediate states. In addition to denoising, we have accomplished translation between paired SAR and IR, as well as SAR and EO images, based on the diffusion Schrödinger bridge framework. These tasks are crucial because they enable the fusion of complementary information from different sensor modalities, improving overall situational awareness and enhancing the quality of remote sensing applications. Remarkably, with only 500 pairs of data, yielding results comparable to those of pix2pix trained on a significantly larger dataset.

4.2 Diffusion Model and Schrödinger Bridge

A diffusion model is a class of generative models that simulates the process of gradually adding noise to data until it becomes indistinguishable from random noise, and then learns to reverse this process to generate data from noise [78]. The concept is insmured by the physical process of diffusion, where particles move from areas of higher concentration to lower concentration, leading to an equilibrium state. In the context of machine learning, diffusion models are trained through a sequence of forward steps that progressively corrupt the data with noise, followed by a learned reverse process that aims to reconstruct the original data from the noisy state. The reverse process is tymucally achieved through a neural network that is trained to predict earlier, less noisy states of the data given its current state. Diffusion models have shown remarkable success in generating high-quality, diverse samples in various domains such as images, audio, and text, distinguishing themselves by their ability to model complex data distributions and produce outputs with high fidelity and variation [80, 81].

The Schrödinger Bridge (SB), dating back to Schrödinger [83] and revisited by Léonard [84], is conceptualized as an entropy-regularized variant of the classical optimal transport problem. This framework encompasses the subsequent stochastic differential equations (SDEs):

$$dX_t = [f_t + \beta_t \nabla \log v(X_t, t)]\, dt + \sqrt{\beta_t} dW_t \tag{4.1}$$

$$d\tilde{X}_t = \left[\tilde{f}_t - \beta_t \nabla \log u(\tilde{X}_t, t)\right] dt + \sqrt{\beta_t} d\tilde{W}_t \tag{4.2}$$

where f is the drift term, W is the Wiener process, β is a constant, $X_0 \sim p_A$ and $X_1 \sim p_B$ are distributed according to the boundary conditions in two discrete domains. The potentials v and u, belonging to the space $C^{1,2}(\mathbb{R}^d \times [0,1])$, are dynamic entities governed by the ensuing coupled partial differential equations (PDEs). Here, the superscript $1,2$ indicates that the functions have continuous first-order derivatives with respect to time and continuous second-order derivatives with respect to spatial variables.

$$\frac{\partial v(x,t)}{\partial t} = -\nabla \cdot (f_t v) + \frac{1}{2}\beta_t \Delta v \tag{4.3}$$

$$\frac{\partial u(x,t)}{\partial t} = -\nabla \cdot (\tilde{f}_t u) + \frac{1}{2}\beta_t \Delta u \tag{4.4}$$

subject to the initial and terminal distribution conditions:

$$v(x,0) = p_A(x), \quad u(x,1) = p_B(x) \tag{4.5}$$

In the formulation eqs. (4.1) to (4.5), the path probability measures induced by the SDEs in (4.3) and (4.4) coincide almost surely, reminiscent of the equivalence established

in the earlier SDEs. Consequently, the marginal probability densities, hereinafter denoted by $q(\cdot,t)$, are correspondingly equivalent.

The insights from existing diffusion models, score-based generative models, and the Schrödinger Bridge problem will inform the development of xDSBMIT. By integrating the stability and interpretability features of the Schrödinger Bridge with the robust image generation capabilities of diffusion models, our approach aims to enhance multimodal image translation. The related works provide foundational principles and techniques that are crucial in formulating and optimizing the Explainable Diffusion Model via Schrödinger Bridge in Multimodal Image Translation framework, which we detail in the subsequent method section.

4.3 xDSBMIT Method

We begin by introducing the fundamental principles underlying our approach. The xDSB leverages the probabilistic nature of diffusion models to transition between image distributions. Specifically, xDSB models the distribution of paired images in the source and target domains, ensuring that essential features of the source images are preserved in the translation process. This is achieved through the Schrödinger Bridge problem, which formulates a continuous path between two probability distributions. The path minimizes the Kullback-Leibler (KL) divergence between the distributions, leading to an optimal transport solution.

To formalize this, let $\mu \in \mathscr{P}_{N+1}$ represent the distribution sequence of diffusion paths, with $\mu_0 = p_A$ and $\mu_N = p_B$ indicating the source and target distributions, respectively. The objective is to find μ^* that satisfies:

$$\mu^* = \arg\min_{\mu}\{\mathrm{KL}(\mu\|\mu_{\mathrm{ref}}) : \mu \in \mathscr{P}_{N+1}, \mu_0 = p_A, \mu_N = p_B\}. \tag{4.6}$$

In multimodal contexts, the distribution of paired images involves diverse modalities [85, 86], such as SAR and IR, ensuring the translation preserves the modality-specific features.

The xDSB training algorithm follows the iterative proportional fitting procedure (IPF) [87] to refine the distribution sequence μ. The goal is to adjust μ iteratively until convergence, ensuring that $\mu_0 \approx p_A$ and $\mu_N \approx p_B$. The IPF updates are given by:

$$\mu^{2n+1} = \arg\min_{\mu}\{\mathrm{KL}(\mu\|\mu^{2n}) : \mu \in \mathscr{P}_{N+1}, \mu_N = p_B\} \tag{4.7}$$

$$\mu^{2n+2} = \arg\min_{\mu}\{\mathrm{KL}(\mu\|\mu^{2n+1}) : \mu \in \mathscr{P}_{N+1}, \mu_0 = p_A\}. \tag{4.8}$$

Each iteration alternates between optimizing the distribution at the source and target ends, gradually refining the transition path to minimize the overall divergence. This method accounts for the multimodal nature of image distributions by considering the unique characteristics of each modality in the optimization process.

For a static version of the problem, we consider the entropy-regularized optimal transport, which links the Schrödinger Bridge problem with traditional optimal transport theory. The SB approach ensures the convergence of the distribution sequence by balancing the entropy terms with the transport cost:

$$\mu^*_{\text{static}} = \arg\min_{\mu}\left\{E_\mu\left[\|x_0 - x_N\|^2\right] - 2\sigma^2 H(\mu) : \mu \in \mathscr{P}_2, \mu_0 = p_A, \mu_N = p_B\right\}. \tag{4.9}$$

The Diffusion Schrödinger Bridge (DSB) combines the dynamic aspects of diffusion processes with the optimal transport properties of the Schrödinger Bridge. In DSB, the forward and backward transition probabilities are updated iteratively to ensure

convergence towards the equilibrium state [88]. The forward and backward passes are given by:

$$p_{t+1|t}(x_{t+1}|x_t) = N(x_{t+1}; x_t + \gamma_t f_t(x_t), 2\gamma_t I), \tag{4.10}$$

$$q_{t|t+1}(x_t|x_{t+1}) = N(x_t; x_{t+1} + \gamma_t b_{t+1}(x_{t+1}), 2\gamma_t I). \tag{4.11}$$

where γ_t is a constant and I is the identity matrix. The training loss functions for DSB are defined to minimize the discrepancies between the forward and backward transitions:

$$L^B_{t+1} = E_{(x_{t+1},x_t)\sim p^n_{t+1,t}} \left[\| B^n_{t+1}(x_{t+1}) - (x_{t+1} + F^n_t(x_t) - F^n_t(x_{t+1})) \|^2 \right], \tag{4.12}$$

$$L^F_{t+1} = E_{(x_t,x_{t+1})\sim q^n_{t,t+1}} \left[\| F^n_t(x_t) - (x_t + B^n_{t+1}(x_{t+1}) - B^n_{t+1}(x)) \|^2 \right]. \tag{4.13}$$

where B and F are two learnable neural networks. The diffusion process of DSB optimizes towards a static goal.

The Explainable Diffusion Model via Schrödinger Bridge (xDSB) integrates diffusion models with the Schrödinger Bridge framework to enhance the stability and interpretability of multimodal image translation. Diffusion models simulate the gradual addition of noise and learn to reverse this process through denoising steps, generating high-quality images. Score-based generative models iteratively refine samples using the score function, the gradient of the log probability density. The Schrödinger Bridge provides an optimal transport solution by creating a continuous path between two probability distributions, minimizing the Kullback-Leibler divergence. In xDSB, iterative

proportional fitting (IPF) refines the distribution sequence to ensure convergence towards an optimal transport path [89]. The model combines static entropy-regularized optimal transport with dynamic diffusion processes, optimizing transitions using neural networks. This approach ensures equilibrium states, enhancing the quality and efficiency of image translations across different modalities. The algorithm design is as follows:

Algorithm 1 Training the xDSB Model

Input: $p_A(\cdot)$ and $p_B(\cdot|X_0)$ datasets
Initialization: Initialize μ and model parameters θ
repeat
    Sample $t \sim \mathscr{U}([0,1])$
    Sample $X_0 \sim p_A(X_0)$, $X_1 \sim p_B(X_1|X_0)$
    Compute $X_t \sim q(X_t|X_0,X_1)$ according to the Schrödinger Bridge formulation
    Update μ using iterative proportional fitting (IPF)
    Perform gradient descent step on $\varepsilon(X_t,t;\theta)$
until convergence =0

Algorithm 2 Generating Images with the Trained xDSB Model

Input: $X_N \sim p_B(X_N)$, trained $\varepsilon(\cdot,\cdot;\theta)$
for $n = N$ to 1 **do**
    Predict X_0^{ε} using $\varepsilon(X_n,t_n;\theta)$
    Sample $X_{n-1} \sim p(X_{n-1}|X_0^{\varepsilon},X_n)$ according to the trained model
end for
Output: X_0 =0

CHAPTER FIVE

TRANSTRANS:A REPRESENTATION-LEVEL TRANSLATION

5.1 Multimodal Sentiment Analysis

The advent of digital media has led to an explosion in the availability of multimodal data, which includes a combination of audio, text, and visual information [90]. The rich and diverse data opens new avenues for research in fields such as sentiment analysis, where understanding human sentiment through different modalities can enhance user experiences, detect emotional well-being, and predict consumer behavior [91]. Traditional sentiment analysis models primarily focused on a single modality, usually text [92]. However, relying solely on text can be limiting, as it often fails to capture the full spectrum of human emotions. For instance, intonation and pitch in audio, or facial expressions in video, provide crucial context that can significantly influence the interpretation of sentiment [93, 94]. Multimodal sentiment analysis leverages the complementary information provided by different modalities to enhance the accuracy and robustness of sentiment predictions [55], and has consequently emerged as a powerful approach by integrating multiple data sources to achieve a more comprehensive understanding of sentiment [95].

Recent advances in deep learning, particularly with Transformer architectures, have further revolutionized the field of multimodal sentiment analysis [17]. Transformers, introduced by Vaswani et al., have demonstrated remarkable success in various domains, including natural language processing (NLP) and computer vision [17]. Their ability to model long-range dependencies and handle different types of input data makes them particularly well-suited for multimodal tasks [96]. For example, the Multimodal Transformer (MulT) by Tsai et al. utilizes cross-modal attention to effectively capture

interactions between audio, text, and visual modalities, significantly improving sentiment analysis performance [55].

Despite these advancements, the robustness of multimodal models in the presence of missing or incomplete data is a critical challenge remains [97]. Real-world applications often encounter scenarios where one or more modalities may be unavailable or corrupted. This missing data can severely impact the performance of multimodal framework, as they rely on the complementary information from all modalities [97]. Current Transformer-based models, while effective in handling multimodal data, struggle to maintain robustness when faced with missing modalities.

Translation-based model can predict and compensate for missing data, ensuring reliable sentiment analysis even when some modalities are absent [47–49, 53]. However, these methods inadequately harness the inherent strengths of individual modalities and employ an excessively intricate translation framework. To address this research gap, we propose a novel Transformer-based translation-driven multimodal sentiment analysis system named TransTrans. Our framework integrates the strengths of Transformer architectures with a translation-based approach to enhance robustness against missing modalities, and different pre-trained models were applied to extract features specific to each modality.

5.2 Transformer-based and Translation-based Method

Transformer models, known for their ability to handle long-range dependencies, have become fundamental in multimodal sentiment analysis. Their adaptability across tasks allows for effective fusion of different modalities, such as audio, text, and visual data. The Multimodal Transformer (MulT) by Tsai et al. [55] and other models like MISA model by Hazarika et al. [98] have demonstrated significant improvements in sentiment analysis by employing cross-modal attention mechanisms to capture interactions between

modalities. These advancements underline the potential of transformer-based models in achieving state-of-the-art performance across various benchmarks. Additionally, Yu et al. [54] presented a hierarchical Transformer model that captures both intra-modal and inter-modal interactions for sentiment analysis, demonstrating significant improvements over traditional methods.

More recently, Wang and Liu proposed a cross-modal Transformer architecture for sentiment analysis, which uses a dual-stream approach to process and fuse multimodal data [99]. This model outperformed existing methods by effectively capturing the interactions between different modalities. Similarly, Wang and He introduced a multimodal sentiment analysis framework based on a hybrid Transformer architecture, combining cross-modal and intra-modal attention mechanisms to achieve state-of-the-art performance [100]. However, these models didn't consider the scenarios of missing or unreliable modalities, which causes performance deterioration under such scenarios. Translation-based approaches have gained significant attention in multimodal sentiment analysis due to their ability to handle missing or noisy data by translating features across different modalities, thereby enhancing model robustness. Liang et al. utilized translation mechanisms within multimodal transformer networks, translating modalities into a shared space to facilitate sentiment analysis in video-grounded dialogue systems [47]. This approach simplifies the fusion process and improves robustness, but may lead to information loss when dealing with significant modality differences, affecting sentiment prediction precision. Tsai et al. proposed a factorized multimodal representation approach that translates information across modalities to enhance sentiment-related features [48]. While this method captures interdependencies effectively, it might struggle with complex multimodal interactions, potentially compromising predictive performance.

Pham et al. introduced a cyclic translation mechanism to learn robust joint representations by cyclically translating features between modalities [49]. This method

maintains high performance even with incomplete data, but the potential introduction of redundant information can increase training complexity. Additionally, the MTMSA model presented by Liu et al. specifically addresses the challenge of uncertain missing modalities by translating visual and audio data into text, utilizing a transformer-based network to ensure robust sentiment predictions even under incomplete data conditions [53].

Our model differentiates itself from the aforementioned translation-based models by leveraging the unique strengths of each modality through modality-specific pre-trained models, integrated within a streamlined Transformer architecture. Unlike existing methods, it avoids complex translation mechanisms, instead directly aligning features across modalities, reducing information loss and computational complexity while maintaining robustness and accuracy, even with incomplete data.

5.3 TransTrans Framework

The proposed framework, TransTrans, consists of three core components: modality-specific feature extraction, translation and concatenation, and sentiment prediction. For feature extraction, the framework uses pre-trained models tailored to each modality, ensuring optimal feature representation for audio, text, and visual data. The extracted features are then passed through a translation mechanism to handle potential missing modalities by predicting the absent data and combining it with the available modalities. This approach ensures robust sentiment prediction, even when some modality data is incomplete, enhancing the overall reliability of the model. Fig. 5.1 illustrates the overall architecture of our system.

We utilize three state-of-the-art models for feature extraction from audio, text, and visual data:

- **CLAP** (Contrastive Language-Audio Pretraining): A pre-trained model designed for extracting high-level audio features [101].

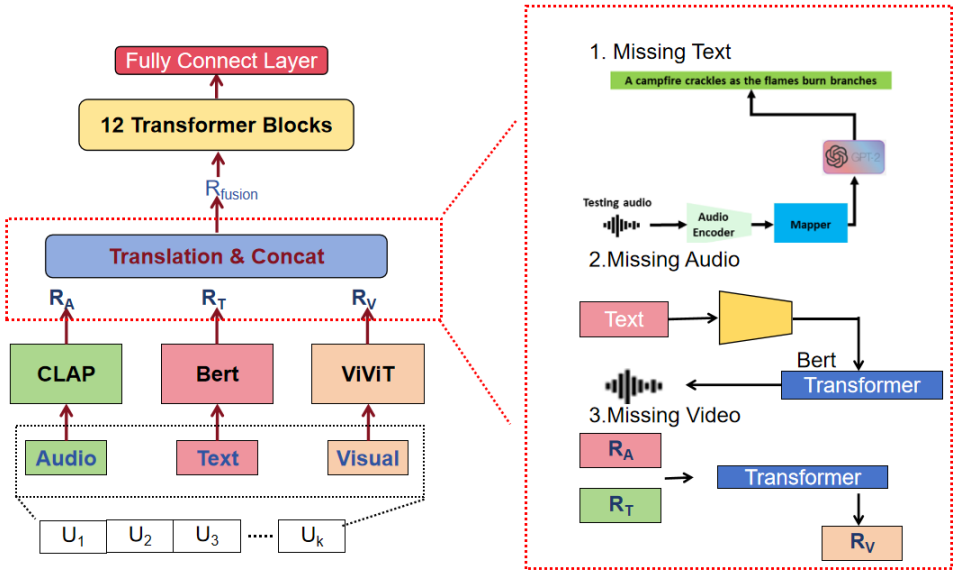

Figure 5.1: The Architecture of TransTrans.

- **BERT** (Bidirectional Encoder Representations from Transformers): A widely-used model for capturing contextual information from text [102].
- **ViViT** (Video Vision Transformer): A Transformer-based model tailored for visual data analysis [103].

Each model processes its respective modality, generating feature representations $\mathbf{R}_A$, $\mathbf{R}_T$, and $\mathbf{R}_V$.

Let $\mathbf{U}_A$, $\mathbf{U}_T$, and $\mathbf{U}_V$ denote the input data for audio, text, and visual data, respectively. The modality-specific feature extraction can be formulated as follows:

$$\mathbf{R}_A = \text{CLAP}(\mathbf{U}_A)$$

$$\mathbf{R}_T = \text{BERT}(\mathbf{U}_T)$$

$$\mathbf{R}_V = \text{ViViT}(\mathbf{U}_V)$$

5.4 Translation and Prediction

To handle cases where one or more modalities are missing, our framework implements specific strategies for each scenario, as illustrated on the right side of Fig. 5.1.

5.4.1 Missing Text

In the case of missing text, we do not require additional training. Instead, we employ GPT-2 to perform cross-modal translation from audio to text. The audio features $\mathbf{R}_A$ are directly passed through a GPT-2 model that generates the corresponding textual features $\hat{\mathbf{R}}_T$. This predicted text representation is then used alongside the available modalities.

5.4.2 Missing Audio

For missing audio, the available text features $\mathbf{R}_T$ are first extracted using BERT. These text features are then passed through four Transformer blocks to predict the missing audio features $\hat{\mathbf{R}}_A$. The training of these blocks is guided by a reconstruction loss, defined as:

$$\mathscr{L}_{\text{audio}} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{R}_A - \hat{\mathbf{R}}_A\|^2$$

This loss ensures that the predicted audio features closely match the original audio features when available.

5.4.3 Missing Video

Similarly, when the video modality is missing, the model uses both available audio and text features $\mathbf{R}_A$ and $\mathbf{R}_T$. These features are concatenated and passed through a separate set of Transformer blocks to predict the missing visual features $\hat{\mathbf{R}}_V$. The reconstruction loss for the visual modality is defined analogously:

$$\mathscr{L}_{\text{video}} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{R}_V - \hat{\mathbf{R}}_V\|^2$$

This approach ensures that the model can robustly handle cases where one or more modalities are absent by effectively reconstructing the missing modality features.

5.4.4 Sentiment Prediction

After obtaining the fused representation $\mathbf{R}_{\text{fusion}}$, which is a concatenation of $\mathbf{R}_A$, $\mathbf{R}_T$, and $\mathbf{R}_V$ or their predicted version, the model proceeds to sentiment prediction using a series of Transformer blocks. Through extensive testing, we determined that utilizing 12 Transformer blocks yields the best performance. These blocks are trained to predict sentiment using a combination of cross-entropy loss and mean squared error loss.

For classification tasks, the cross-entropy loss $\mathscr{L}_{\text{CE}}$ is defined as:

$$\mathscr{L}_{\text{CE}} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i)$$

where C is the number of sentiment classes, y_i is the true label, and $\hat{y}_i$ is the predicted probability.

For regression tasks, particularly when predicting continuous sentiment scores, the mean squared error loss $\mathscr{L}_{\text{MSE}}$ is used:

$$\mathscr{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

These losses guide the training of the Transformer blocks, ensuring that the model can accurately predict sentiment based on the fused multimodal representation.

CHAPTER SIX

RESULTS

6.1 Image-to-Image Translation

In our study, we utilize a comprehensive collection of datasets spanning different imaging modalities to support our research on multi-modal image translation which is one key application of end-to-end translation. Synthetic Aperture Radar (SAR) images capture radar signal representations of the Earth's surface, providing valuable data under conditions where optical sensors might fail, such as in adverse weather. Infrared (IR) images offer critical insights into thermal properties of the landscape, important for environmental monitoring. Electro-optical (EO) sensors contribute images in the visible spectrum, including both RGB and grayscale images, which are predominantly used in computer vision applications due to their detailed representation of visible light information. Lastly, the RGB images provide high-resolution, color imagery of agricultural and urban landscapes. The data in our experiments are sourced from the UNICORN dataset and the PBVS 2024 public competition. The UNICORN dataset comprises paired SAR and EO data, while the dataset provided in the PBVS 2024 open competition includes paired SAR, IR, and RGB data, as illustrated in Fig. 6.1. These datasets form a robust foundation for exploring and enhancing techniques in image translation across various modalities.

Using an NVIDIA RTX8000 GPU, xDSBMIT demonstrated significant efficiency compared to DDPM. During the training phase, xDSBMIT had notably lower runtimes, as shown in Fig 6.2, it took 4.3 hours to train on 1000 data points, whereas DDPM required 7.5 hours. Similarly, in the testing phase, as shown in Fig 6.3, xDSBMIT completed testing in 2.5 hours for 1000 data points, compared to 3.8 hours for DDPM. Overall,

Figure 6.1: Paired Images: SAR, IR and RGB

xDSBMIT exhibited superior efficiency in both training and testing, highlighting its advantages for real-world applications.

The SAR2IR task utilizes the dataset provided in the PBVS 2024 open competition. As shown in Fig. 2, we demonstrated the feasibility of translating Synthetic Aperture Radar (SAR) images into Infrared (IR) imagery. The translated IR images effectively reconstructed the primary contours and structural details present in the original SAR data, as evidenced by the image sequence in the middle column of Fig. 6.4. However, despite these promising results, the translated images exhibited diminished brightness and were unable to capture some finer details compared to the original IR images. These observations suggest that while the approach is effective in capturing major features, further refinement is needed to enhance the detail fidelity and brightness levels of the translated images, thus improving their utility for practical applications in remote sensing.

In the subsequent SAR2EO translation experiment, notable advancements were achieved using a relatively modest dataset of only 500 training images from UNICORN

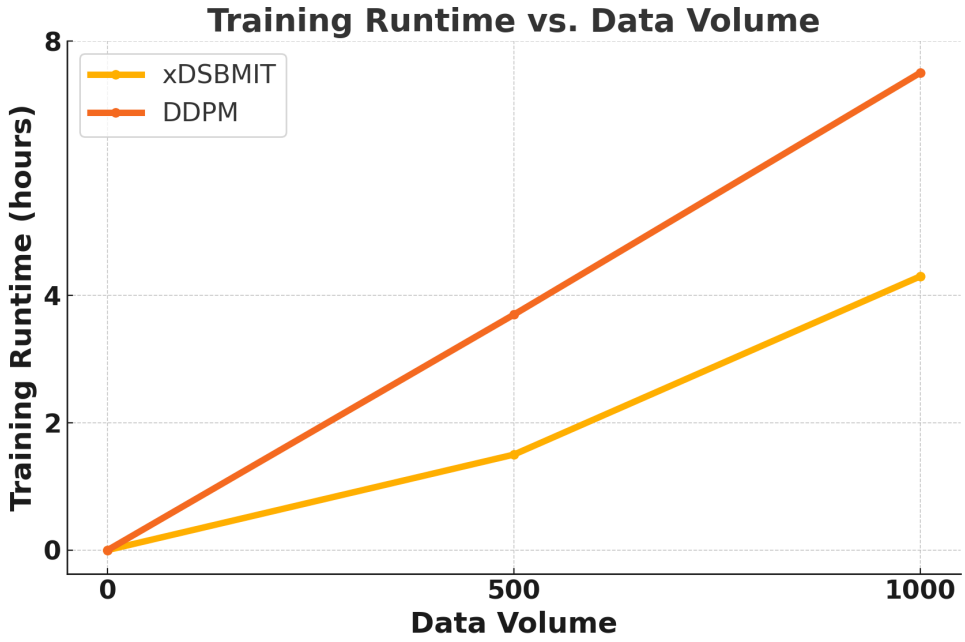

Figure 6.2: Training Runtime

dataset. Despite the limited data, our approach surpassed the performance benchmarks set by well-established frameworks such as pix2pix, pix2pixHD, and traditional GAN models. The results indicate a significant improvement not only in the accuracy of the translated EO images but also in the clarity and color fidelity. This breakthrough demonstrates the potential of our model to efficiently learn and generalize from sparse datasets, outperforming existing methods in both qualitative and quantitative evaluations. The successful application with minimal training data underscores our model's robustness and efficiency, suggesting it as a highly effective tool for enhancing EO image generation in remote sensing technologies. Table 6.1 provides a performance comparison of different image translation methods for SAR2EO. Our model, referred to as EDSB-500, exhibits superior performance across both LPIPS and FID metrics. Specifically, EDSB-500 achieved an LPIPS score of 0.35 and an FID score of 0.10, outperforming the GAN-500,

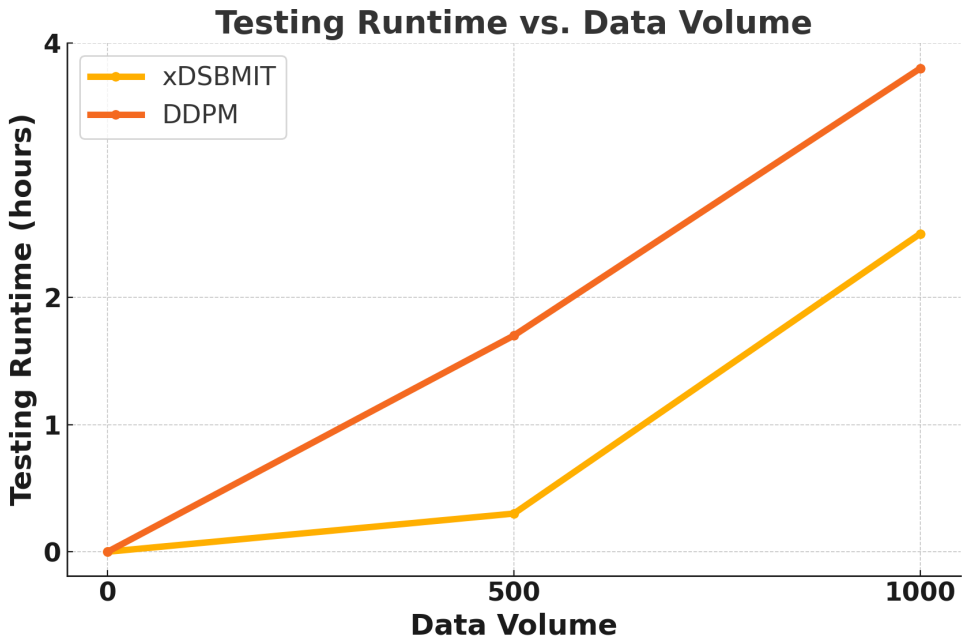

Figure 6.3: Testing Runtime

pix2pix-500, and pix2pixHD-500 models, which recorded higher LPIPS and FID scores. LPIPS (Learned Perceptual Image Patch Similarity) measures the perceptual similarity between generated and real images. The LPIPS is constructed based on the VGG-16 [104] architecture. Lower LPIPS scores indicate higher perceptual quality, as the generated images are closer to the real ones in terms of human visual perception. In our experiment, the EDSB-500 model achieved the lowest LPIPS score, suggesting that it produces more perceptually accurate images compared to the other models. FID (Fréchet Inception Distance) evaluates the quality of generated images by comparing the distributions of real and generated image features extracted by a pre-trained Inception network [105]. Lower FID scores indicate that the generated images have a distribution closer to the real images, thus reflecting higher quality. The EDSB-500 model achieved the lowest FID score, indicating a significant improvement in image quality and fidelity over the other models.

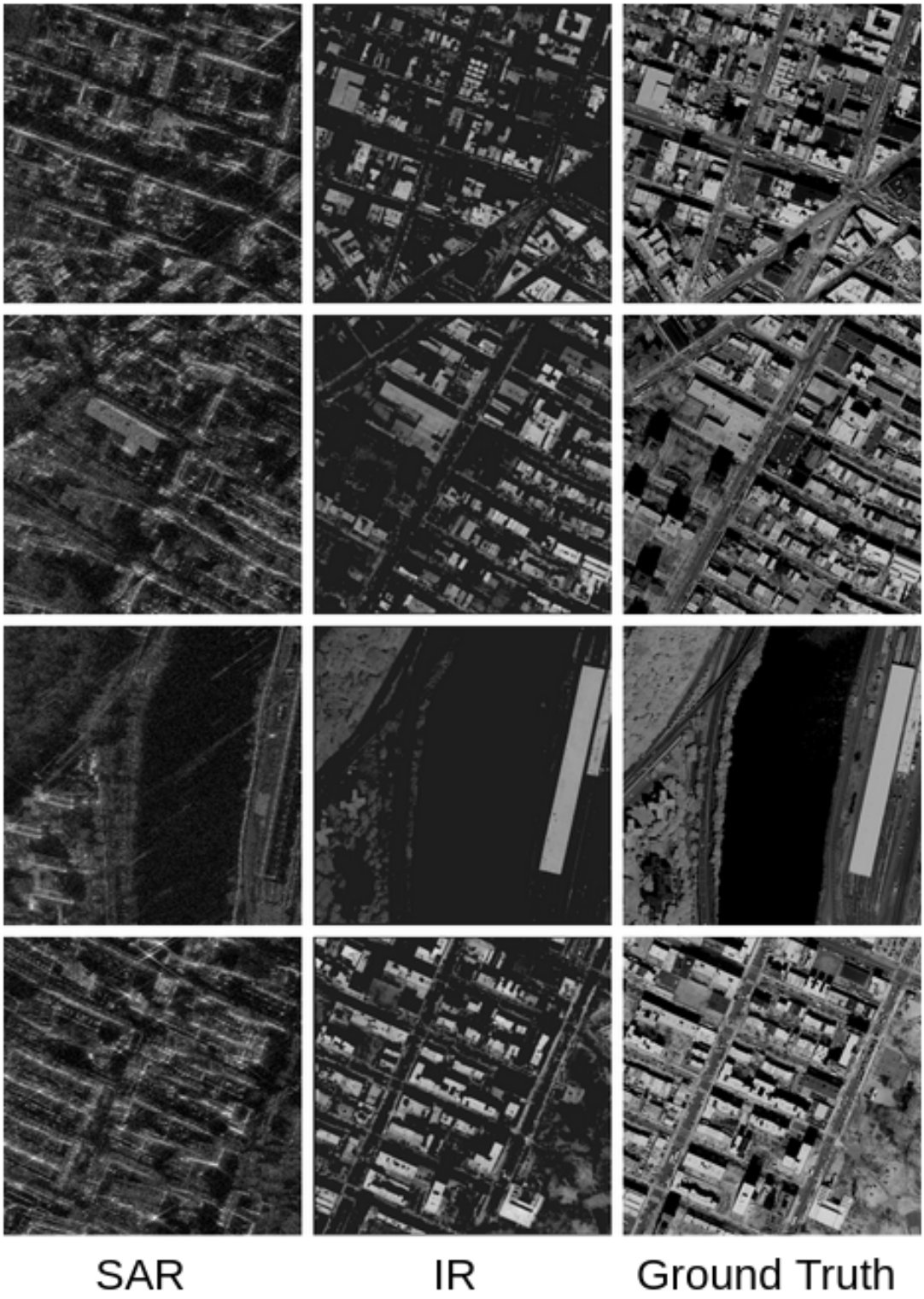

Figure 6.4: SAR2IR translation. Left: SAR. Middle: IR generated via translation. Right: Ground Truth of IR.

These results highlight the effectiveness of our model in generating high-quality EO images from SAR data, even with a limited training dataset.

Table 6.1: Performance comparison of different image translation methods for SAR2EO

	LPIPS	FID
GAN-500	0.52	0.44
pix2pix-500	0.48	0.27
pix2pixHD-500	0.45	0.18
xDSBMIT-500	**0.35**	**0.10**

6.2 Multimodal Sentiment Analysis

To show the performance of reprensentation-level translation, we evaluate the TransTrans framework on two benchmark datasets of multimodal sentiment analysis, CMU-MOSI and CMU-MOSEI. Each dataset contains synchronized audio, text, and visual data with sentiment annotations. The evaluation metrics include Mean Absolute Error (MAE), Correlation (Corr), two-class accuracy (Acc-2), F1 score, and seven-class accuracy (Acc-7).

Firstly, to find out the original capability, we conducted a series of experiments to compare the performance of TransTrans with other state-of-the-art multimodal sentiment analysis models, including TFN [106], ICCN [107], MulT [55], BERT [102], Self-MM [45], and Modified TEASEL [108], with all three modalities presented. The results are summarized in Table 1.

As shown in Table 1, TransTrans achieves the best performance across all metrics, including the lowest MAE of 0.628, the highest correlation of 0.833, the highest two-class

Table 6.2: Performance comparison with state-of-the-art sentiment analysis models on CMU-MOSI.

Model	MAE	Corr	Acc-2	F1	Acc-7
BERT	0.739	0.782	85.20	85.20	-
Self-MM	0.713	0.798	85.98	85.94	-
TFN	0.901	0.698	80.28	80.77	34.94
ICCN	0.860	0.710	83.00	83.21	39.23
MulT	0.871	0.698	83.02	82.80	40.10
TEASEL	0.632	0.812	86.92	85.31	45.52
TransTrans	**0.628**	**0.833**	**87.24**	**85.4**	**46.68**

accuracy of 87.24%, the high F1 score of 85.4%, and the highest seven-class accuracy of 46.68%. These results demonstrate the effectiveness of our translation-based approach in enhancing the robustness and accuracy of multimodal sentiment analysis.

Table 6.3: Comparison of translation-based methods in missing modality experiments on CMU-MOSI.

Model	Acc-0.2	Acc-0.5
AE	78.03	69.30
MCTN	77.21	69.98
MTMSA	83.85	**79.16**
TransTrans	**83.93**	78.34

Then it's the experiments of missing modalities. We simulate the scenarios by randomly deleting specific level of original data from each modality and compare TransTrans with other translation-based sentiment analysis methods. There are three experiments corresponding to missing audio, missing text and missing video, individually. The results are the average of the three experiments. Table 2 presents the results on the

CMU-MOSI dataset. Acc-0.2 corresponds to the condition where 20% of a single modality is missing, while Acc-0.5 represents the scenario where 50% of a modality is missing.

From the results, it is evident that TransTrans model achieves the best or close to the best performance in both scenarios, obtaining 83.93% for Acc-0.2 and 78.34% for Acc-0.5. This demonstrates that TransTrans is highly effective at handling missing modalities, maintaining robust accuracy even when 50% of the modality data is absent. MTMSA obtains 83.85% for Acc-0.2 and 79.16% for Acc-0.5, which is comparable to the performance of TransTrans. Other models, such as AE [109] and MCTN [49], show significantly lower performance, particularly under the Acc-0.5 condition. This highlights the robust performance of TransTrans model in dealing with substantial modality loss.

Table 6.4: Ablation study on translation mechanism in CMU-MOSI

Modality Combination	Model	Accuracy
Video+Audio	Self-MM	0.783
	TransTrans	0.832
Video+Text	Self-MM	0.830
	TransTrans	0.847
Text+Audio	Self-MM	0.649
	TransTrans	0.825

We also performed an ablation study to evaluate the impact of the translation mechanism on our model's performance. The results are summarized in Table 3. We tested the accuracy of our model under different modality combinations when one modality is missing.

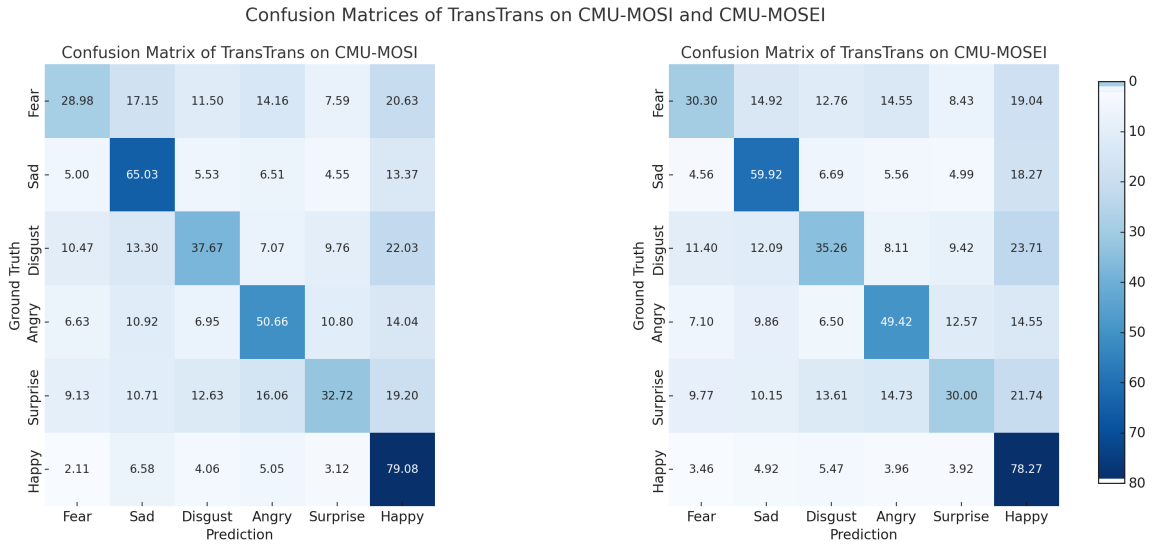

Figure 6.5: Confusion Matrices of TransTrans on CMU-MOSI and CMU-MOSEI

The results in Table 3 show that our model consistently outperforms the baseline Self-MM [45] across all modality combinations. For instance, when visual modality is missing, our model achieves an accuracy of 0.832 compared to 0.783 with the Self-MM model. Similarly, when the text modality is missing, our model achieves an accuracy of 0.847 compared to 0.830 with the Self-MM model. These improvements highlight the effectiveness of our translation mechanism in handling missing modalities.

These experimental results validate the effectiveness of TransTrans framework in multimodal sentiment analysis. By incorporating a translation mechanism, TransTrans not only improves the robustness of the system against missing modalities but also enhances the overall performance across various metrics.

Fig. 6.5 shows the confusion matrices for the TransTrans model on the CMU-MOSI and CMU-MOSEI datasets, providing a comprehensive view of the model's classification performance across various emotions. The model demonstrates good accuracy in predicting "Sad" and "Happy" emotions, achieving 65.03% and 79.08% on

CMU-MOSI, and 59.92% and 78.27% on CMU-MOSEI, respectively. However, it is evident that “Disgust” and “Fear” are more challenging to classify, as these emotions are frequently misclassified. The model shows a significant tendency to confuse “Disgust” with “Angry” and struggles with distinguishing “Fear” from other emotions, indicating that these two categories share overlapping features that the model finds difficult to separate. This consistent challenge across both datasets highlights the need for further refinement in handling subtle and complex emotional expressions, particularly for “Disgust” and “Fear”.

CHAPTER SEVEN

CONCLUSION

7.1 Overview of Thesis

This thesis has explored translation-based multimodal learning, focusing on two complementary models: xDSBMIT and TransTrans. The xDSBMIT framework integrates the Diffusion Schrödinger Bridge model for image translation tasks, specifically addressing the challenges of translating SAR to EO and IR images. This approach provided significant improvements in translation quality, even with limited datasets, demonstrating stability and interpretability through its unique integration of diffusion processes. TransTrans, on the other hand, tackled the issues of multimodal sentiment analysis, leveraging translation-driven learning to handle missing modalities using Transformer-based architecture. By focusing on representation-level translation, TransTrans demonstrated resilience and accuracy in sentiment prediction, outperforming existing models, especially in scenarios with incomplete data.

The thesis provided a detailed experimental evaluation of these models, showing that both xDSBMIT and TransTrans contribute substantially to advancing the state-of-the-art in their respective application domains. xDSBMIT excels in interpreting and generating high-quality image translations from different sensor modalities, while TransTrans effectively manages multimodal data for sentiment analysis, even under missing data conditions. Together, these models reflect the power and potential of translation-based multimodal learning in diverse fields, ranging from remote sensing to social media analysis.

7.2 Future Directions

While this thesis has addressed several challenges in translation-based multimodal learning, there are numerous opportunities for further exploration. One promising direction is enhancing the scalability of xDSBMIT to accommodate larger datasets and different sensor modalities. Expanding its application to other domains, such as medical imaging, climate analysis, or multimodal Unmanned Aerial Vehicle (UAV) data, could further demonstrate its versatility and robustness. Additionally, optimizing the diffusion process for computational efficiency may allow for faster training and inference times, which is crucial for real-time applications.

For TransTrans, future work could focus on enhancing the model's ability to handle more complex missing modality scenarios, particularly involving dynamic data like video or streaming information. Integrating additional modalities, such as physiological signals like EEG and eye movement, could provide richer multimodal data and improve sentiment analysis accuracy. Moreover, exploring unsupervised or semi-supervised approaches could address the issue of data scarcity and reduce the reliance on labeled datasets, making the model more adaptable to real-world applications.

Overall, the findings of this thesis lay the foundation for future advancements in translation-based multimodal learning, and the proposed directions aim to further improve robustness, applicability, and scalability in practical scenarios.

REFERENCES

[1] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223, 2016.

[2] R. Tylecek and R. Sára, "Spatial pattern templates for recognition of objects with regular structure," in *Proceedings of the German Conference on Pattern Recognition*, pp. 364–374, 2013.

[3] T. Sattler, Q. Dai, M. Fritz, and L. Van Gool, "Aachen day-night: A dataset for large-scale scene recognition across different lighting conditions," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 540–547, 2017.

[4] C. Leong, T. Rovito, O. Mendoza-Schrock, C. Menart, J. Bowser, L. Moore, S. Scarborough, M. Minardi, and D. Hascher, "Unified coincident optical and radar for recognition (unicorn) 2008 dataset," 2008. Dataset.

[5] W. R. Tan, C. S. Chan, H. Aguirre, and K. Tanaka, "Improved artgan for conditional synthesis of natural image and artwork," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 394–409, 2019.

[6] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2641–2649, 2015.

[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

[8] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2018.

[9] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audioset: An ontology and human-labeled dataset for audio events," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, 2017.

[10] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv preprint arXiv:1606.06259*, 2016.

[11] A. Zadeh, P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Cmu-mosei: A multimodal dataset for sentiment analysis and emotion recognition," *arXiv preprint arXiv:1803.09457*, 2018.

[12] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 335–338, 2008.

[13] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[14] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 127–133, 2003.

[15] N. Kalchbrenner and P. Blunsom, "Recurrent continuous translation models," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu, and S. Bethard, eds.), (Seattle, Washington, USA), pp. 1700–1709, Association for Computational Linguistics, Oct. 2013.

[16] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, p. 3104–3112, 2014.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, vol. 30, 2017.

[18] D. Bahdanau, "Neural machine translation by jointly learning to align and translate," *arXiv preprint. https://arxiv.org/abs/1409.0473*, 2014.

[19] Y. Pang, J. Lin, T. Qin, and Z. Chen, "Image-to-image translation: Methods and applications," *IEEE Transactions on Multimedia*, vol. 24, pp. 3859–3881, 2022.

[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[21] M. Mirza, "Conditional generative adversarial nets," *arXiv preprint. https://arxiv.org/abs/1411.1784*, 2014.

[22] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

[23] D. P. Kingma, "Auto-encoding variational bayes," *arXiv preprint, https://arxiv.org/abs/1312.6114*, 2013.

[24] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8798–8807, 2018.

[25] L. A. Gatys, "A neural algorithm of artistic style," *arXiv preprint, https://arxiv.org/abs/1508.06576*, 2015.

[26] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2016.

[27] D. Nie, R. Trullo, J. Lian, L. Wang, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, "Medical image synthesis with deep convolutional adversarial networks," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 12, pp. 2720–2730, 2018.

[28] Z. Shi, P. Mettes, G. Zheng, and C. Snoek, "Frequency-supervised mr-to-ct image synthesis," in *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections: First Workshop, DGM4MICCAI 2021, and First Workshop, DALI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 1*, pp. 3–13, 2021.

[29] X. Shao and W. Zhang, "Spatchgan: A statistical feature based discriminator for unsupervised image-to-image translation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6546–6555, 2021.

[30] L. Wang, Y. Chae, and K.-J. Yoon, "Dual transfer learning for event-based end-task prediction via pluggable event to image translation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2135–2145, 2021.

[31] S. Du, J. Yu, G. Xie, R. Lu, P. Li, Z. Cai, and K. Lu, "Sar2eo: A high-resolution image translation framework with denoising enhancement," in *Australasian Joint Conference on Artificial Intelligence*, pp. 91–102, Springer, 2023.

[32] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pp. 382–398, Springer, 2016.

[33] S. Li, Z. Tao, K. Li, and Y. R. Fu, "Visual to text: Survey of image and video captioning," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 3, pp. 297–312, 2019.

[34] M. Żelaszczyk and J. Mańdziuk, "Cross-modal text and visual generation: A systematic review. part 1: Image to text," *Information Fusion*, vol. 93, pp. 302–329, 2023.

[35] X. He and L. Deng, "Deep learning for image-to-text generation: A technical overview," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 109–116, 2017.

[36] S. Indurthi, M. A. Zaidi, N. Kumar Lakumarapu, B. Lee, H. Han, S. Ahn, S. Kim, C. Kim, and I. Hwang, "Task aware multi-task learning for speech to text tasks," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7723–7727, 2021.

[37] G. I. Gállego, I. Tsiamas, C. Escolano, J. A. R. Fonollosa, and M. R. Costa-jussà, "End-to-end speech translation with pre-trained models and adapters: Upc at iwslt 2021," 2021.

[38] X. Wang, T. Qiao, J. Zhu, A. Hanjalic, and O. Scharenborg, "Generating images from spoken descriptions," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, p. 850–865, jan 2021.

[39] H. Ning, X. Zheng, Y. Yuan, and X. Lu, "Audio description from image by modal translation network," *Neurocomputing*, vol. 423, p. 124–134, Jan. 2021.

[40] G. Parmar, T. Park, S. Narasimhan, and J.-Y. Zhu, "One-step image translation with text-to-image models," *arXiv preprint, https://arxiv.org/abs/2403.12036*, 2024.

[41] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," in *International Conference on Machine Learning*, 2011.

[42] Y. Bengio, A. C. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 1798–1828, 2012.

[43] H. Yu, L. Gui, M. Madaio, A. Ogan, J. Cassell, and L.-P. Morency, "Temporally selective attention model for social and affective state recognition in multimedia content," in *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, p. 1743–1751, Association for Computing Machinery, 2017.

[44] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly fine-tuning" bert-like" self supervised models to improve multimodal speech emotion recognition," *arXiv preprint, https://arxiv.org/abs/2008.06682*, 2020.

[45] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," 2021.

[46] S. Lai, X. Hu, H. Xu, Z. Ren, and Z. Liu, "Multimodal sentiment analysis: A survey," *Displays*, p. 102563, 2023.

[47] H. Le, D. Sahoo, N. Chen, and S. Hoi, "Multimodal transformer networks for end-to-end video-grounded dialogue systems," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (A. Korhonen, D. Traum, and L. Màrquez, eds.), (Florence, Italy), pp. 5612–5623, Association for Computational Linguistics, July 2019.

[48] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," *arXiv preprint, https://arxiv.org/abs/1806.06176*, 2018.

[49] T. Pham, T. Tran, D. Phung, and S. Venkatesh, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 6892–6899, 2019.

[50] C. Shang, A. Palmer, J. Sun, K.-S. Chen, J. Lu, and J. Bi, "Vigan: Missing view imputation with generative adversarial networks," in *2017 IEEE International Conference on Big Data (Big Data)*, pp. 766–775, 2017.

[51] C. Zhang, Y. Cui, Z. Han, J. T. Zhou, H. Fu, and Q. Hu, "Deep partial multi-view learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 5, pp. 2402–2415, 2020.

[52] T. Zhou, S. Canu, P. Vera, and S. Ruan, "Feature-enhanced generation and multi-modality fusion based deep neural network for brain tumor segmentation with missing mr modalities," *Neurocomputing*, vol. 466, pp. 102–112, 2021.

[53] Z. Liu, B. Zhou, D. Chu, Y. Sun, and L. Meng, "Modality translation-based multimodal sentiment analysis under uncertain missing modalities," *Information Fusion*, vol. 101, p. 101973, 2024.

[54] J. Yu, L. Pan, R. Song, C. Wang, and L. Zhang, "Hierarchical transformer for multimodal sentiment analysis," *IEEE Transactions on Multimedia*, vol. 24, pp. 2267–2279, 2021.

[55] Y.-H. H. Tsai, S. Bai, M. Yamada, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, 2019.

[56] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[57] Y. Shi, V. De Bortoli, A. Campbell, and A. Doucet, "Diffusion schrödinger bridge matching," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[58] G.-H. Liu, A. Vahdat, D.-A. Huang, E. A. Theodorou, W. Nie, and A. Anandkumar, "I^2sb: Image-to-image schr\" odinger bridge," *arXiv preprint, https://arxiv.org/abs/2302.05872*, 2023.

[59] Z. Chen, G. He, K. Zheng, X. Tan, and J. Zhu, "Schrodinger bridges beat diffusion models on text-to-speech synthesis," *arXiv preprintm https://arxiv.org/abs/2312.03491*, 2023.

[60] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988, 2022.

[61] B. Bischke, P. Helber, F. Koenig, D. Borth, and A. Dengel, "Overcoming missing and incomplete modalities with generative adversarial networks for building footprint segmentation," in *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*, pp. 1–6, IEEE, 2018.

[62] M. Hamghalam, A. F. Frangi, B. Lei, and A. L. Simpson, "Modality completion via gaussian process prior variational autoencoders for multi-modal glioma segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 442–452, 2021.

[63] S. Li, Z. Tao, K. Li, and Y. Fu, "Visual to text: Survey of image and video captioning," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 3, no. 4, pp. 297–312, 2019.

[64] T. Zhou, S. Canu, P. Vera, and S. Ruan, "Latent correlation representation learning for brain tumor segmentation with missing mri modalities," *IEEE Transactions on Image Processing*, vol. 30, pp. 4263–4274, 2021.

[65] J. Sun, X. Zhang, S. Han, Y.-P. Ruan, and T. Li, "Redcore: Relative advantage aware cross-modal representation learning for missing modalities with imbalanced missing rates," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 13, pp. 15173–15182, 2024.

[66] K. R. Park, H. J. Lee, and J. U. Kim, "Learning trimodal relation for audio-visual question answering with missing modality," *arXiv preprint, https://arxiv.org/abs/2407.16171*, 2024.

[67] D. Kim and T. Kim, "Missing modality prediction for unpaired multimodal learning via joint embedding of unimodal models," 2024.

[68] Z. Guo, T. Jin, and Z. Zhao, "Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1726–1736, 2024.

[69] X. Lin, S. Wang, R. Cai, Y. Liu, Y. Fu, W. Tang, Z. Yu, and A. Kot, "Suppress and rebalance: Towards generalized multi-modal face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 211–221, 2024.

[70] S. Liu, M. Gao, V. John, Z. Liu, and E. Blasch, "Deep learning thermal image translation for night vision perception," *ACM Transactions on Intelligent Systems and Technology*, vol. 12, pp. 1–18, Dec 2020.

[71] S. Low, O. Nina, A. D. Sappa, E. Blasch, and N. Inkawhich, "Multi-modal aerial view image challenge: Translation from synthetic aperture radar to electro-optical domain results-pbvs 2023," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 515–522, 2023.

[72] A. Savakis, A. M. N. Taufique, and N. Nagananda, "Advances in domain adaptation for aerial imagery," in *Handbook of Dynamic Data Driven Applications Systems* (F. Darema, E. P. Blasch, S. Ravela, and A. J. Aved, eds.), Cham: Springer, 2023.

[73] Y. Zheng, E. Blasch, and Z. Liu, *Multispectral Image Fusion and Colorization*. SPIE Press, 2018.

[74] E. P. Blasch, F. Darema, S. Ravela, and A. J. Aved, eds., *Handbook of Dynamic Data Driven Applications Systems*, vol. 1. Springer, 2nd ed., 2022.

[75] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017.

[76] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *International Conference on Learning Representations*, 2019.

[77] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[78] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proceedings of the International Conference on Machine Learning*, 2015.

[79] M. Özbey *et al.*, "Unsupervised medical image translation with adversarial diffusion models," *IEEE Transactions on Medical Imaging*, vol. 42, no. 12, pp. 3524–3539, 2023.

[80] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Advances in Neural Information Processing Systems*, 2021.

[81] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, 2020.

[82] V. De Bortoli, J. Thornton, J. Heng, and A. Doucet, "Diffusion schrödinger bridge with applications to score-based generative modeling," in *Advances in Neural Information Processing Systems*, vol. 34, pp. 17695–17709, 2021.

[83] E. Schrödinger, "über die umkehrung der naturgesetze," *Sitzungsberichte der Preussischen Akademie der Wissenschaften Physikalisch-Mathematische Klasse*, 1932.

[84] C. Léonard, "A survey of the schrödinger problem and some of its connections with optimal transport," *Discrete and Continuous Dynamical Systems*, 2013.

[85] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 172–189, 2018.

[86] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *International Conference on Machine Learning*, pp. 1857–1865, 2017.

[87] W. E. Deming and F. F. Stephan, "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known," *Annals of Mathematical Statistics*, vol. 11, no. 4, pp. 427–444, 1940.

[88] Z. Tang, T. Hang, S. Gu, D. Chen, and B. Guo, "Simplified diffusion schrödinger bridge," *arXiv preprint, https://arxiv.org/abs/2403.14623*, 2024.

[89] L. Rüschendorf and W. Thomsen, "Note on the schrödinger equation and i-projections," *Statistics & Probability Letters*, vol. 17, no. 5, pp. 369–375, 1993.

[90] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[91] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[92] A. Balahur, P. Rosso, and A. Montoyo, "Sentiment analysis in social media texts," in *WASSA 2012*, p. 110, 2012.

[93] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102–107, 2017.

[94] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proceedings of the 13th International Conference on Multimodal Interfaces*, pp. 169–176, 2011.

[95] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *Information Fusion*, vol. 50, pp. 69–79, 2019.

[96] J. Lin, J. Chen, Q. Liu, and Z. Li, "Transformers in vision: A survey," *ACM Computing Surveys*, vol. 55, no. 4, pp. 1–36, 2022.

[97] B. Liu, Z. Huang, S.-L. Yuan, Z. Yang, and Y. Wang, "Multimodal machine learning: A survey," *ACM Computing Surveys*, vol. 54, no. 4, pp. 1–25, 2021.

[98] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and -specific representations for multimodal sentiment analysis," *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[99] D. Wang, S. Liu, Q. Wang, Y. Tian, L. He, and X. Gao, "Cross-modal enhancement network for multimodal sentiment analysis," *IEEE Transactions on Multimedia*, vol. 25, pp. 4909–4921, 2023.

[100] Y. Wang, J. He, D. Wang, Q. Wang, B. Wan, and X. Luo, "Multimodal transformer with adaptive modality weighting for multimodal sentiment analysis," *Neurocomputing*, vol. 572, p. 127181, 2024.

[101] T. Zhao, M. Kong, T. Liang, Q. Zhu, K. Kuang, and F. Wu, "CLAP: Contrastive language-audio pre-training model for multi-modal sentiment analysis," in *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, p. 622–626, 2023.

[102] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1, p. 2, 2019.

[103] H. Akbari, W. Yin, T. Darrell, and S. Mandt, "Vivit: A video vision transformer," *IEEE Transactions on Image Processing*, 2021.

[104] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[105] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *CoRR*, 2015.

[106] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114, 2017.

[107] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 8992–8999, 2020.

[108] M. Arjmand, M. J. Dousti, and H. Moradi, "Teasel: a transformer-based speech-prefixed language model," *arXiv preprint, https://arxiv.org/abs/2109.05522*, 2021.

[109] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," in *Proceedings of ICML Workshop on Unsupervised and Transfer Learning* (I. Guyon, G. Dror, V. Lemaire, G. Taylor, and D. Silver, eds.), vol. 27 of *Proceedings of Machine Learning Research*, (Bellevue, Washington, USA), pp. 37–49, PMLR, 02 Jul 2012.