Semantic Change Driven Generative Semantic Communication Framework

Wanting Yang[a], Zehui Xiong[a], Hongyang Du[b], Yanli Yuan[c], Tony Q. S. Quek[a]
[a]Information Systems Technology and Design Pillar, Singapore University of Technology and Design
[b] School of Computer Science and Engineering, Nanyang Technological University
[c] School of Cyberspace Science and Technology, Beijing Institute of Technology
{wanting_yang, zehui_xiong, tonyquek}@sutd.edu.sg, hongyang001@e.ntu.edu.sg, yanliyuan@bit.edu.cn

2024 IEEE Wireless Communications and Networking Conference (WCNC) | 979-8-3503-0358-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/WCNC57260.2024.10571010

***Abstract*—The burgeoning generative artificial intelligence technology offers novel insights into the development of semantic communication (SemCom) frameworks. These frameworks hold the potential to address the challenges associated with the black-box nature inherent in existing end-to-end training manner for the existing SemCom framework, as well as deterioration of the user experience caused by the inevitable error floor in deep learning-based SemCom. In this paper, we focus on the widespread remote monitoring scenario, and propose a semantic change driven generative SemCom framework. Therein, the semantic encoder and semantic decoder can be optimized independently. Specifically, we develop a modular semantic encoder with value of information based semantic sampling function. In addition, we propose a conditional denoising diffusion probabilistic mode-assisted semantic decoder that relies on received semantic information from the source, namely, the semantic map, and the local static scene information to remotely regenerate scenes. Moreover, we demonstrate the effectiveness of the proposed semantic encoder and decoder as well as the considerable potential in reducing energy consumption through simulation based on the realistic $\mathcal{F}$ composite channel fading model. The code is available at https://github.com/wty2011jl/SCDGSC.git.**

***Index Terms*—Conditional DDPM, semantic sampling, generative AI, remote monitoring, value of information**

I. Introduction

As the human society progresses toward the remote management and automation, a high-capacity system ensuring reliability and cost/power-efficiency will emerge as an imperative requirement [1]. To achieve optimal effectiveness and the sustainability against this background, semantic communication (SemCom) has garnered considerable attention as a promising solution. Exploiting the intelligence of the communicating parties, SemCom shifts the focus from "how" to transmit to "what" to transmit [2], so as to boost network performance by reducing the data required to be transmitted.

Specifically, the research endeavors in SemCom primarily concentrate on how to extract absolutely required information for recovering the meaning, i.e., the design of the semantic encoder and semantic decoder. The existing SemCom framework can be broadly divided into three categories based on the adopted semantic extraction methods, i.e., deep learning (DL) based SemCom, reinforcement learning (RL) based SemCom, and knowledge base (KB) assisted SemCom [3]. Among them, DL-based SemCom has emerged as the prevailing choice for a diverse spectrum of communication tasks [4]. Nevertheless, the intrinsic limitations of DL give rise to two inevitable challenges with SemCom. Firstly, the requirement for a differentiable loss function places constraints on the selection of metrics guiding the training process, thereby diminishing its adaptability to a diverse range of tasks. Secondly, the inherent error floor in DL results in sub-optimal communication performance, even under ideal channel conditions.

To address the first challenge, RL-based SemCom has been proposed to allow for the incorporation of more semantic metrics into the training process [5]. However, it should be noted that RL-based SemCom is only applicable to sequence-generation tasks due to the recurrent nature of the actor network. In addition, both DL-based and RL-based semantic communication exhibit a complete end-to-end black-box nature, which limits these SemCom frameworks' social acceptance and practicality in complex network environments [3]. In comparison, KB-based SemCom excels in terms of explainability. Nevertheless, due to the high computational complexity involved in constructing the KB itself, the application of KB-assisted SemCom to real-time, on-demand tasks is challenging [6].

To address the aforementioned issues, generative artificial intelligence, particularly the recently prominent conditional denoising diffusion probabilistic model (DDPM), offers novel insights for the advancement of semantic encoders and decoders. In real life, most communication scenarios do not necessitate perfect recovery at the bit or pixel level; rather, they require the retrieved data to closely approximate reality while ensuring the preservation of complete semantic information. This ensures a favorable quality of experience (QoE). For example, in traffic flow monitoring or parking space surveillance, the system remains indifferent to particulars like the color of the vehicle. Instead, it exclusively focuses on the vehicle's location, necessitating a high level of image clarity. To this end, we introduce the conditional DDPM into the SemCom framework, where the sender only needs to transmit essential semantic information, expertly extracted by a customised semantic encoder, as a prompt to the receiver with the DDPM assisted semantic decoder. The receiver, in

The research is supported by the National Research Foundation (NRF) and Infocomm Media Development Authority under the Future Communications Research Development Programme (FCP). The research is supported by the Ministry of Education, Singapore, under its SUTD Kickstarter Initiative (SKI-20210204). The research is also supported by the Ministry of Education, Singapore, under its-SMU-SUTD-Joint Grant (22-SIS-SMU-048).

979-8-3503-0358-2/24/$31.00 ©2024 IEEE
Authorized licensed use limited to: Singapore University of Technology & Design. Downloaded on January 05,2026 at 08:43:04 UTC from IEEE Xplore. Restrictions apply.

turn, utilizes this prompt to steer purposeful generation to fulfil the task of semantic decoding. Thanks to the noteworthy accomplishments in a plethora of real-world generation tasks, especially the photo-realistic images generation, the DDPM model based semantic decoder exhibits the capability to alleviate the deterioration of QoE resulting from inevitable error floors inherent in DL-based SemCom framework.

Taking the above into consideration, we propose a novel generative Semcom framework explicitly tailored for the widespread remote monitoring scenario for the first time, called semantic change driven generative semantic communication (SCDGSC) framework. In contrast to the three prevailing SemCom frameworks, all of which adhere to the end-to-end training paradigm, the generative SemCom distinguishes itself by affording the opportunity for independent design and optimization of the semantic encoder and semantic decoder, thereby enhancing the explainability of semantic information. The specific contributions are as follows.

- We have developed a modular semantic encoder endowed with semantic sampling capabilities. In this design, we introduce a more sophisticated semantic criterion for sampling beyond age, which we refer to as value of information (VoI). Given that the primary concern of the receivers predominantly revolves around changes that exert influence on subsequent tasks, the measurement of VoI concurrently encompasses the semantic change degree of the observed scene and the age of information (AoI), where the changes irrelevant to the task are omitted.
- We have designed a DDPM-assisted semantic decoder, which exclusively relies on the semantic information conveyed by the source, specifically, the semantic map, for the purpose of remote scene generation. Moreover, in order to generate a close-to-real remote scene, an image of static information of the remote scene is also used as one of the inputs to the semantic decoder. Since it is only updated by the source to the destination when static information alterations, such as changes in weather conditions or the time of day, the corresponding communication overhead is negligible.
- We have conducted training on the models of the target segmentation module and the DDPM-based scene generation module within semantic encoder and semantic decoder, respectively. The adopted dataset is generated from CDnet2014 [7]. Subsequently, we have evaluated the efficiency of these models and have provided substantial evidence of the framework's considerable potential in reducing energy consumption through simulation based on the realistic $\mathcal{F}$ composite channel fading model.

II. System Model

In this work, we focus on a single-source and single-server remote status update system. Specifically, the source in the considered scenario is an embedded vision sensor with limited memory and computing power, which is responsible for monitoring a certain scene, generating a series of visual samples, and updating the destination on the sampled image timely via the wireless and wired transmission. The destination is a remote server, which is for reproducing the remote monitoring scene in real time for situational awareness, location tracking, control etc. [8], based on the received samples and the built-in predictive estimation algorithm, such as Kalman Filter and future frame prediction.

In sharp contrast to the conventional communication dedicated to the optimization of transmission process, (where the sampled data flow is usually modelled as a stationary stochastic process or it is assumed that the source adopts periodic sampling), in this work, the optimization of sampling process is also factored into the communication process. The embedded source can semantically sample and extract most pivotal information to promise a significant reduction in transmission burden, thus saving considerable resources while guaranteeing communications performance. The details of the implementation for the proposed tailored SemCom framework are presented and studied in Section III.

In addition, we believe that wireless transmission is a bottleneck in the communication process. In this sense, we mainly analyze and evaluate the wireless transmission performance in this work. Considering the combined effects of multi-path and shadowing on the practical transmission, we adopt the $\mathcal{F}$ composite fading model to characterize the stochastic wireless channel [9]. We denote the instantaneous channel gain by $\tilde{g}$. The probability density function of $\tilde{g}$ is expressed by [9]

$$f\left(\tilde{g}\right)=\frac{m^{m}(m_{s}-1)^{m_{s}}\bar{g}^{m_{s}}\tilde{g}^{m-1}}{B\left(m,m_{s}\right)\left[m\tilde{g}+\left(m_{s}-1\right)\bar{g}\right]^{m+m_{s}}}, \tag{1}$$

where m, m_s represents the number of clusters of multipath, shadowing shape, respectively, and $\bar{g}$ is corresponding average channel gain, i.e., $\bar{g}=\mathbb{E}\left[\tilde{g}\right]$. Moreover, $B\left(\cdot,\cdot\right)$ denotes the beta function [9]. In this work, we focus on the design of the semantic encoder and decoder. Without loss of generality, we assume perfect capacity achieving coding in this work. It is assumed that in order to cope with stochastic fading, the transmitter of the embedded vision sensor adopts a power control technique. We denote the decoding threshold of signal-to-noise ratio by Θ. In this sense, the achievable transmission rate is expressed by

$$R=W\log\left(1+\Theta\right), \tag{2}$$

where W is the allocated bandwidth. Then, the instantaneous transmit power is expressed as

$$\tilde{p}=\frac{\Theta\sigma^{2}}{\tilde{g}}. \tag{3}$$

Since the stochastic fading can be treated as independently and identically distributed (i.i.d.) among transmission time intervals, the average transmit power over the time can be expressed by

$$\begin{aligned}\bar{p}&=\mathbb{E}_{\tilde{g}}\left[\tilde{p}\right]=\int_{0}^{\infty}\frac{\Theta\sigma^{2}}{\tilde{g}}f\left(\tilde{g}\right)d\tilde{g}\\&=\Theta\sigma^{2}\int_{0}^{\infty}\tilde{g}^{-1}f\left(\tilde{g}\right)d\tilde{g}\\&=\Theta\sigma^{2}\mathbb{E}\left[\tilde{g}^{-1}\right].\end{aligned} \tag{4}$$

According to (1), with the aid of [10, eq. (3.194.3)], the n^{th}

Authorized licensed use limited to: Singapore University of Technology & Design. Downloaded on January 05,2026 at 08:43:04 UTC from IEEE Xplore. Restrictions apply.

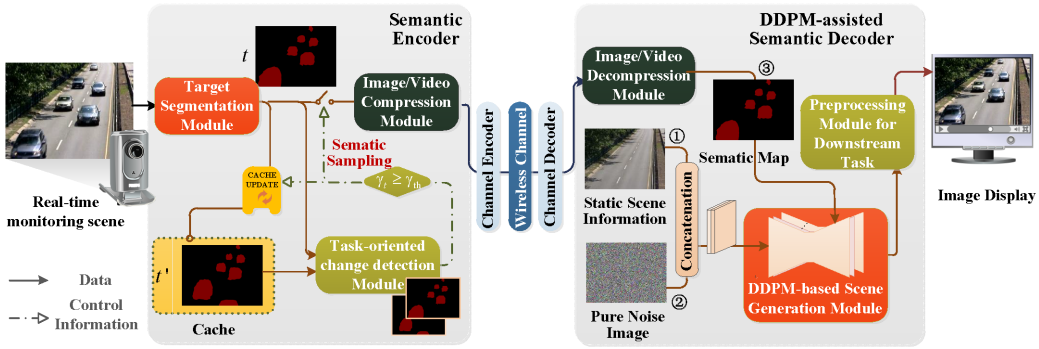


Fig. 1: Semantic change driven generative SemCom framework.

moment of $\tilde{g}$ can be derived as

$$\mathbb{E}\left[\tilde{g}^n\right] = \frac{(m_s-1)^n \bar{g}^n \Gamma(m+n)\,\Gamma(m_s-n)}{m^n \Gamma(m)\,\Gamma(m_s)}, \tag{5}$$

where $\Gamma(\cdot)$ represents the gamma function. Substituting the case of $n=-1$ in (5) into (4), we can obtain the final expression of $\bar{p}$ as below,

$$\bar{p} = \frac{\Theta\sigma^2 m \Gamma(m-1)\,\Gamma(m_s+1)}{(m_s-1)\,\bar{g}\,\Gamma(m)\,\Gamma(m_s)}. \tag{6}$$

III. SCDGSC: Framework and Key Components

A. Framework Overview

In this work, we develop a novel SCDGSC framework, as illustrated in Fig. 1. In contrast to the existing SemCom model characterized by black-box nature of the end-to-end training manner, a divide-and-conquer approach is adopted in the proposed framework. Moreover, compared to the source encoder in the conventional communication, which only serves the role of video or image compression, two additional modules, *target segmentation* and *VoI-based semantic sampling*, are integrated in the deigned semantic encoder. Accordingly, the semantic decoder is also integrated two modules, *DDPM-based scene generation* and *preprocessing module for downstream task*, which correspond to target segmentation module and VoI-based sampling module, respectively.

It is widely recognized that, in the remote monitoring system, only the information about changes is of concern, especially the changes in mission-related objectives. With this in mind, the captured scene is first fed into the target segmentation module to extract the location and contour information of the objectives of interest and simultaneously timestamped with t. The output of the target segmentation module is hereinafter referred to as the *semantic map*, which is denoted by $\mathbf{s}$. For tracking changes in the scene, the source maintains a cache to store the last semantic map updated to the destination, the timestamp of which is denoted by t'. To judge the value of the newly captured scene, both the semantic maps $\mathbf{s}_t$ and $\mathbf{s}_{t'}$ are fed into the VoI-based semantic sampling module. Different from the AoI-oriented real-time tracking systems, the module for semantic sampling not only takes the age $|t-t'|$ into account, but the semantic changes in the observed scene. We assume that a VoI threshold (denoted by $\gamma_{\rm th}$) is set in the system. If γ_t is less than $\gamma_{\rm th}$, the current semantic map $\mathbf{s}_t$ is discarded directly at the source. Otherwise, the semantic map $\mathbf{s}_{t'}$ is replaced by $\mathbf{s}_t$ for the next sampling judgement. At the same time, the semantic map $\mathbf{s}_t$ is fed into the compression model and then sent to the destination as an update sample.

As opposed to the process of semantic encoding, the received compressed updated samples after the channel decoder are firstly fed into the image/video decompression module. Then, the recovered semantic maps are taken as one of the three inputs of DDPM-based scene generation module to guide the generation of the real-time remote scene. Moreover, for synthesizing a seamless and realistic scene, the static scene information of the remote scene also acts as an input of the generation module, which is denoted by $\mathbf{r}$. In addition, the third input is a pure noise image, which is the outcome of the Gaussian-based forward process of DDPM model itself. At last, the generated scene images are input into the preprocesing module for the downstream task. Take video surveillance reconstruction as an example. The generated images as well as their timestamp information can be input into a frame prediction module, which can present users with the illusion of the real-time transmission for the remote scene, by leveraging the frame prediction techniques. The details of the key component design can be found in the next subsection.

B. Key Component Design

Given the space limitation, only three modules proposed in this work, *target segmentation*, *VoI-based semantic sample*, and *DDPM-based scene generation*, are presented here for the implementation process.

1) Target Segmentation: The design of the target segmentation can be divided into four parts. The first block is called the initial block, which serves primarily to reduce the size of the subsequent feature map by down-sampling the captured scene image. The second part is the backbone, which is employed to extract the semantic information embedded in the original image. Given the limited computing capacity of the embedded vision sensor, we opt for the MobileNetV3 backbone [11] in this work due to its commendable balance between computational efficiency and accuracy. During the semantic extraction, four feature maps can be acquired, two of which are fed into the Lite-R-ASPP segmentation head [11] for the further semantic aggregation and the final target segmentation results. Moreover, to ensure the transmission quality, we add a channel adaptive interpolate block to resize the semantic map before

Authorized licensed use limited to: Singapore University of Technology & Design. Downloaded on January 05,2026 at 08:43:04 UTC from IEEE Xplore. Restrictions apply.

output, which is controlled by a down-sampling parameter. The specific mapping of down-sampling parameters to channel conditions needs further research in particular scenarios.

2) VoI-based Semantic Sampling Module: The VoI metric considered in this work encompasses two aspects. One is the age of the sample updated to the destination, which can be approximated by calculating the difference between the timestamps of the current captured scene and the last updated scene. As stated in Section III-A, it can be expressed by

$$\gamma_t^{\text{AoI}} = |t - t'| . \tag{7}$$

The other is the semantic change degree, which can be obtained by comparing the differences between two semantic maps $\mathbf{s}_t$ and $\mathbf{s}_{t'}$. Since the semantic map contains only the location and contour information of the target of interest to the task, the changes about the irrelevant information, such as times of the day and weather, are self-ignored during the comparison. We denote the total number of the pixels occupied by the task-relevant objectives in semantic maps $\mathbf{s}_t$ and $\mathbf{s}'_t$ by n_t and $n_{t'}$, respectively. Moreover, the number of the pixels in the intersection of the pixel set occupied by the objectives in the two semantic maps is denoted by $n_{tt'}$. Then, the semantic change degree can be expressed by

$$\gamma_t^{\text{change}} = \frac{n_t + n_{t'} - 2n_{tt'}}{n_t + n_{t'}}. \tag{8}$$

From (8), we can see that, if the target regions in the two semantic maps overlap exactly, $\gamma_{\text{change}} = 0$, which means that the scene has not undergone a semantic change in the meantime. In contrast, if the target regions in the two semantic maps are completely separated, $\gamma_{\text{change}} = 1$, which is the maximum value of the semantic change degree. Since in real-world scenarios, the generation at the destination side needs to rely on predictive techniques, the information that the scene has not changed can also facilitate a better grasp of the evolution of the environment. With this in mind, considering the both factors, the complete expression of VoI can be formulated as

$$\gamma_t = \tau_1 \gamma_t^{\text{AoI}} + \tau_2 \gamma_t^{\text{change}}, \tag{9}$$

where parameters τ_1 and τ_2 are used to adjust the sensitivity of the VoI to the semantic change degree and AoI.

3) DDPM-based Scene Generation Model: Inspired of the remarkable success of the DDPM model in a plethora of real-world generation tasks, we involve the conditional DDPM into the SemCom framework as the core of semantic decoder. As discussed in Section III-A, both static scene $\mathbf{r}$ and the real-time semantic map $\mathbf{s}_t$ are taken as the inputs to guide the generation process. Therefore, the conditional DDPM can be treated as a latent variable model of the form $p_\theta(\mathbf{x}_0) := \int p_\theta(\mathbf{x}_{0:N} | \mathbf{r}, \mathbf{s}_t) \, d\mathbf{x}_{0:N}$, where $\mathbf{x}_1, \ldots, \mathbf{x}_T$ are latents with the same dimensionality as the possible generated data $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, and $p_\theta(\mathbf{x}_{0:N} | \mathbf{r}, \mathbf{s}_t) = p(\mathbf{x}_N) \prod_{n=1}^{N} p_\theta(\mathbf{x}_{n-1} | \mathbf{x}_n, \mathbf{r}, \mathbf{s}_t)$. The joint distribution $p_\theta(\mathbf{x}_{0:N} | \mathbf{r}, \mathbf{s}_t)$ is called the reverse process, which can be modelled as a Markov chain with learned Gaussian transitions

$$p_\theta(\mathbf{x}_{n-1} | \mathbf{x}_n, \mathbf{r}, \mathbf{s}_t) = \mathcal{N}\left(\mathbf{x}_{n-1}; \tilde{\mu}_\theta(\mathbf{x}_n, \mathbf{x}_0), \tilde{\beta}_\theta I\right), \tag{10}$$

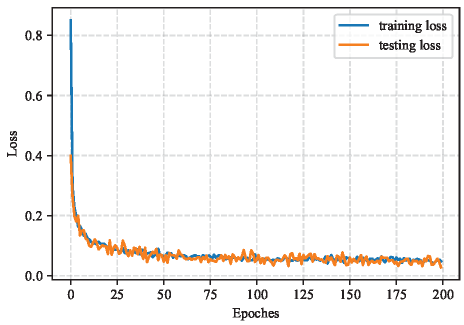


Fig. 2: Loss function for conditional DDPM.

starting at a pure noise image $\mathbf{x}_N \sim \mathcal{N}(\mathbf{x}_N; \mathbf{0}, \mathbf{I})$. To facilitate learning $p_\theta(\mathbf{x}_{0:N} | \mathbf{r}, \mathbf{s}_t)$, an approximate posterior $q(\mathbf{x}_{1:N} | \mathbf{x}_0)$ named forward process is defined and fixed to a Markov chain that progressively adds Gaussian noise into the data under a variance schedule $\beta_1, \ldots, \beta_N$[1]. Given that $\bar{\alpha}_n := \prod_{s=1}^{n} (1 - \beta_s)$, the forward process is expressed by

$$q(\mathbf{x}_n | \mathbf{x}_{n-1}) = \mathcal{N}\left(\mathbf{x}_n; \sqrt{1 - \beta_n}\mathbf{x}_0, \beta_n \mathbf{I}\right), \tag{11}$$

$$q(\mathbf{x}_n | \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_n; \sqrt{\bar{\alpha}_n}\mathbf{x}_0, (1 - \bar{\alpha}_n)\mathbf{I}\right). \tag{12}$$

Moreover, following the properties of the Gaussian distribution, the posteriors of the forward process, $q(\mathbf{x}_{n-1} | \mathbf{x}_n, \mathbf{x}_0)$, also obeys a Gaussian distribution, i.e.,

$$q(\mathbf{x}_{n-1} | \mathbf{x}_n, \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_{n-1}; \tilde{\mu}_n(\mathbf{x}_n, \mathbf{x}_0), \tilde{\beta}_n \mathbf{I}\right). \tag{13}$$

Based on (11) and (12), we have [12]

$$\tilde{\mu}_n(x_n, x_0) = \frac{1}{\sqrt{1 - \beta_n}}\mathbf{x}_n - \frac{\beta_n}{\sqrt{1 - \beta_n}\sqrt{1 - \bar{\alpha}_n}}\epsilon, \tag{14}$$

$$\tilde{\beta}_n = \frac{1 - \bar{\alpha}_{n-1}}{1 - \bar{\alpha}_n}\beta_n. \tag{15}$$

The conditional DDPM is trained to optimize the upper variational bound on negative log likelihood via minimizing the gap between $q(\mathbf{x}_{n-1} | \mathbf{x}_n, \mathbf{x}_0)$ and $p_\theta(\mathbf{x}_{n-1} | \mathbf{x}_n, \mathbf{r}, \mathbf{s}_t)$. According to (15), in our work, the coefficient of variance can be considered as a constant. As such, in this DDPM, it is only required to make $\tilde{\mu}_\theta(\mathbf{x}_n, \mathbf{x}_0)$ as close as possible to $\tilde{\mu}_n(\mathbf{x}_n, \mathbf{x}_0)$. By the observation of (14), an U-Net network $\hat{\epsilon}_\theta(\mathbf{x}_n, \mathbf{r}, \mathbf{s}_t)$ is employed to approximate the noise ϵ generated in each step.

Following the formulation of [12], the simplified denoising loss function can be expressed by

$$\mathcal{L}_d = \mathbb{E}_{n, \mathbf{x}_n, \epsilon}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_n}\mathbf{x}_n + \sqrt{1 - \bar{\alpha}_n}\epsilon, \mathbf{r}, \mathbf{s}_t, n\right)\right\|_2\right]. \tag{16}$$

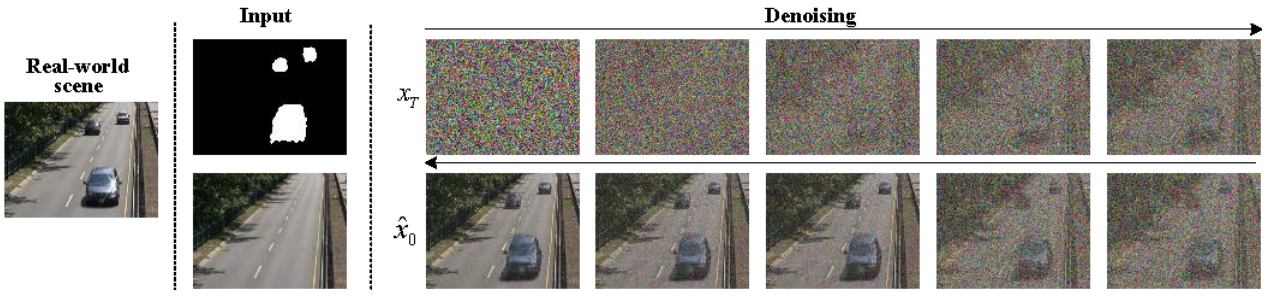
In addition, according to [13], the performance of conditional DDPM can be enhance by gradient of the log probability

[1]In this work, the variance schedule is held constant as hyperparameters.

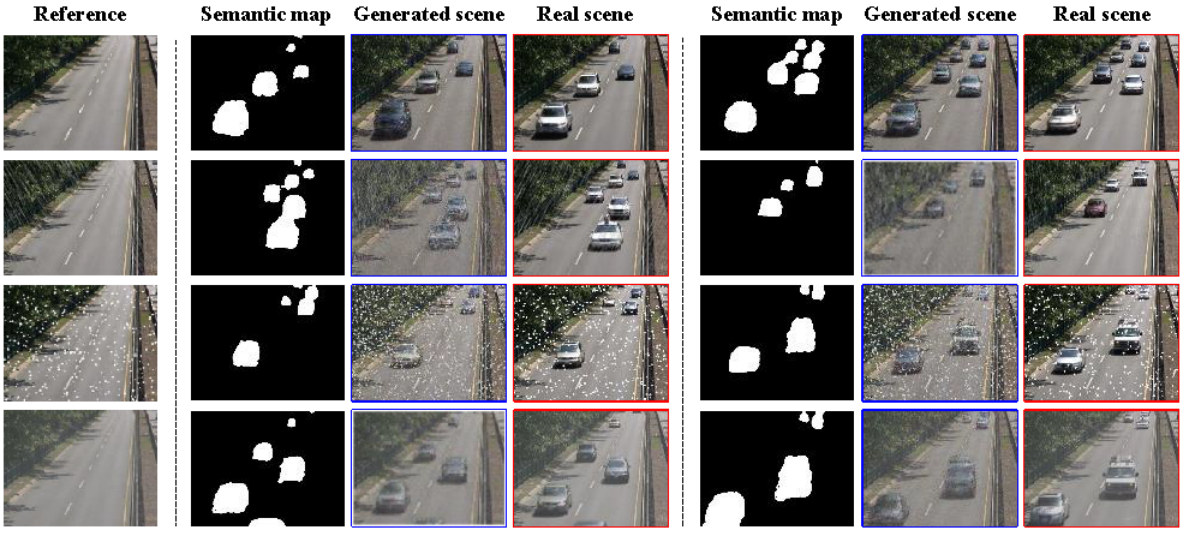
TABLE I: Performance of target segmentation.

Metrics	Value
Average row correct	['99.5','71.9']
Intersection over Union (IoU)	['97.9','65.6']
mean IoU	81.7

Authorized licensed use limited to: Singapore University of Technology & Design. Downloaded on January 05,2026 at 08:43:04 UTC from IEEE Xplore. Restrictions apply.



(a)



(b)

Fig. 3: Visual results. (a) Inference procedure; (b) Performance comparison in different weather conditions.

distribution $\nabla_{\mathbf{x}_n} \log p\left(\mathbf{s}_t \left|\mathbf{x}_n\right.\right)$. In this work, we adopt the classifier-free guidance to implicitly infers the gradient of the log probability [14], as shown in (17). Specifically, the semantic map $\mathbf{s}_t$ is replaced with a null label $\emptyset$ to disentangle the noise estimated under the guidance of semantic map $\epsilon_\theta\left(\mathbf{x}_n \left|\mathbf{r}, \mathbf{s}_t\right.\right)$ from unconditional situation $\epsilon_\theta\left(\mathbf{x}_n \left|\mathbf{r}\right.\right)$.

$$\begin{aligned} &\epsilon_\theta\left(\mathbf{x}_n \left|\mathbf{r}, \mathbf{s}_t\right.\right)-\epsilon_\theta\left(\mathbf{x}_n \left|\mathbf{r}\right.\right) \\ \propto &\nabla_{\mathbf{x}_n} \log p\left(\mathbf{x}_n \left|\mathbf{s}_t, \mathbf{r}\right.\right)-\nabla_{\mathbf{x}_n} \log p\left(\mathbf{x}_n \left|\mathbf{r}\right.\right) \\ \propto &\nabla_{\mathbf{x}_n} \log p\left(\mathbf{s}_t \left|\mathbf{x}_n\right.\right) \end{aligned} \tag{17}$$

Thus, the noise estimation can be performed based on the disentangled component, which can be refined as

$$\hat{\epsilon}_\theta\left(\mathbf{x}_n \left|\mathbf{r}, \mathbf{s}_t\right.\right)=\epsilon_\theta\left(\mathbf{x}_n \left|\mathbf{r}, \mathbf{s}_t\right.\right)+k \cdot\left(\epsilon_\theta\left(\mathbf{x}_n \left|\mathbf{r}, \mathbf{s}_t\right.\right)-\epsilon_\theta\left(\mathbf{x}_n \left|\mathbf{r}\right.\right)\right), \tag{18}$$

where k is the guidance scale, which allows the generated data to follow the semantic map more strictly.

IV. Evaluation

A. Setup

1) Datasets: We select data from the 'baseline' category within the CDNet2014 dataset [7], specifically focusing on road traffic scenarios, as our training and evaluation dataset for target segmentation and conditional DDPM. Specifically, in the training of conditional DDPM, the image of moment"t0" in the CDnet2014 is treated as the reference image and the results of the target segmentation are used as the semantic map. The images at other moments are treated as the labels. We add weather filters for rain, snow and fog to all the images. The training and test sets are divided in a ratio of 8 : 2. The image size is reshaped into $(128, 96)$.

2) Hyperparameters: For the target segmentation, we employ the architecture of Lite R-ASPP and trained on the basis of the pre-training weights obtained by pre-training on COCO [2]. Moreover, we set the number of class as 2. For the conditional DDPM, the structure of employed U-Net network can be referred to in [14]. It comprises a number of channels equal to [64, 64, 128, 128, 256, 256, 512, and 512]. We have $T = 1000$, and a linear variance schedule. Finally, the guidance scale k is set to 4. The batch size is set to 6 and the learning rate is set to $2e-5$.

3) Simulation parameter: For the adopted $\mathcal{F}$ composite fading model, we take fading severity of $m = 6$, shadowing shape $m_s = 6$. Moreover, the average channel gain is treated as the pass loss, which is modeled as $35.3 + 37.6\log_{10}(d)$ in dB. The distance between the vision sensor and the wireless base station is $d = 100$m. The SNR threshold is set to 15 dB.

[2]https://download.pytorch.orgmodelslraspp_mobilenet_v3_large-d234d4ea.pth

Authorized licensed use limited to: Singapore University of Technology & Design. Downloaded on January 05,2026 at 08:43:04 UTC from IEEE Xplore. Restrictions apply.

The bandwidth is $W = 1$ MHz. The noising power is -90 dBm/Hz. Moreover, given the scenario we focus on, we set $\tau_1 = 0$ and $\tau_2 = 1$.

B. Results Analysis

Applying the re-trained model of Lite-R-ASPP [11] to the test set, the performance results are summarized in Table I. Taking the output of the Lite-R-ASPP network as one of the inputs of the conditional DDPM, the training and the testing loss of the DDPM is shown in Fig. 2. The visual results can be found in Fig. 3. Specifically, Fig. 3(a) shows the denoising process. The image in the left column is the real-world scene. The middle column shows the semantic map and the reference image, which can be considered as the prompt for the virtual scene generation. The ten images depicted on the right are captured from the denoising process from step 999 to step 0. By comparing the generated image denoted as $\hat{x}_0$ and the original image, it becomes evident that, while there exist disparities in the specific details and coloration of the vehicles compared to the real scene, the positional alignment of the vehicles remains consistent. This achievement signifies the system's proficiency in preserving essential semantic information, and also allows the framework to be used in the common scenarios that are not concerned with specific details, such as car park space monitoring, traffic flow monitoring, etc. Additionally, as demonstrated in Fig. 3(b), even in the face of diverse weather conditions, the transmission of a semantic map of the real scene suffices. The receiver side adaptively reconstructs a virtual image at the remote location, leveraging the local reference image as a foundation. We adopt the JPEG image compression technique. The datasize of the image of the real scene shown in Fig. 3(a) in different weather and the semantic map are 93 kb, 96 kb, 82 kb, 128 kb and 5 kb, respectively. The comparison of energy consumption of the image transmission is shown in Fig. 4(a). In addition, when we set the VoI thresholds to different values, the comparison of the total energy of data transmission when monitoring the same scene is shown in Fig. 4(b). Simulation results unequivocally underscore the substantial potential of the proposed framework in reducing energy consumption, which contributes to the sustainability of embedded vision sensors.

V. Conclusion

In this paper, we focused on the remote monitoring scenario and proposed a semantic change driven generative semantic communication framework. The distinctiveness of the envisaged framework resided in the design of semantic encoder and decoder, which were bolstered by the advanced semantic segmentation and conditional generative AI techniques, respectively. In contrast to the existing SemCom model characterized by black-box nature of the end-to-end training manner, a divide-and-conquer approach was adopted in the proposed framework. Simulation results demonstrated the effectiveness of the proposed framework and considerable potential for energy savings. In the future, we will systematically investigate semantic sampling optimization within resource constraints and the optimal VoI threshold in this framework, employing diverse performance metrics for downstream tasks.

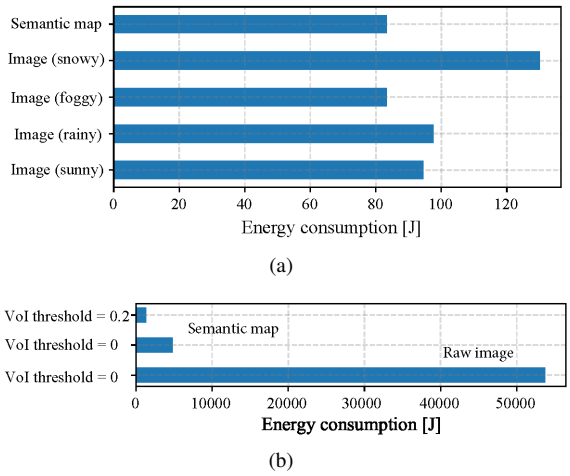


Fig. 4: Comparison of energy consumption. (a) Comparison under different weather conditions; (b) Comparison with different VoI thresholds.

References

[1] S. Iyer, R. Khanai, D. Torse, R. J. Pandya, K. M. Rabie, K. Pai, W. U. Khan, and Z. Fadlullah, "A survey on semantic communications for intelligent wireless networks," *Wireless Personal Communications*, vol. 129, no. 1, pp. 569–611, 2023.

[2] P. Popovski, O. Simeone, F. Boccardi, D. Gündüz, and O. Sahin, "Semantic-effectiveness filtering and control for post-5g wireless connectivity," *Journal of the Indian Institute of Science*, vol. 100, pp. 435–443, 2020.

[3] W. Yang, H. Du, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. S. Shen, and C. Miao, "Semantic communications for future Internet: Fundamentals, applications, and challenges," *IEEE Communications Surveys & Tutorials*, 2022.

[4] Z. Qin, X. Tao, J. Lu, W. Tong, and G. Y. Li, "Semantic communications: Principles and challenges," *arXiv preprint arXiv:2201.01389*, 2021.

[5] K. Lu, Q. Zhou, R. Li, Z. Zhao, X. Chen, J. Wu, and H. Zhang, "Rethinking modern communication from semantic coding to semantic communication," *IEEE Wireless Communications*, vol. 30, no. 1, pp. 158–164, 2022.

[6] W. Yang, X. Chi, L. Zhao, Z. Xiong, and W. Jiang, "Task-driven semantic-aware green cooperative transmission strategy for vehicular networks," *IEEE Transactions on Communications*, 2023.

[7] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, "CDnet 2014: An expanded change detection benchmark dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014, pp. 387–394.

[8] E. Uysal, O. Kaya, A. Ephremides, J. Gross, M. Codreanu, P. Popovski, M. Assaad, G. Liva, A. Munari, B. Soret, T. Soleymani, and K. H. Johansson, "Semantic Communications in Networked Systems: A Data Significance Perspective," *IEEE Network*, vol. 36, no. 4, pp. 233–240, 4 2022.

[9] S. K. Yoo, P. C. Sofotasios, S. L. Cotton, S. Muhaidat, F. J. Lopez-Martinez, J. M. Romero-Jerez, and G. K. Karagiannidis, "A comprehensive analysis of the achievable channel capacity in $\mathcal{F}$ composite fading channels," *IEEE Access*, vol. 7, pp. 34 078–34 094, 2019.

[10] D. Zwillinger and A. Jeffrey, *Table of integrals, series, and products*. Elsevier, 2007.

[11] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.

[12] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[13] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[14] W. Wang, J. Bao, W. Zhou, D. Chen, D. Chen, L. Yuan, and H. Li, "Semantic image synthesis via diffusion models," *arXiv preprint arXiv:2207.00050*, 2022.

Authorized licensed use limited to: Singapore University of Technology & Design. Downloaded on January 05,2026 at 08:43:04 UTC from IEEE Xplore. Restrictions apply.