IEEE Multimedia

CAMUL: Context-Aware Multi-conditional Instance Synthesis for Image Segmentation

Thanh-Danh Nguyen[1,2], Trong-Tai Dam Vu[1,2], Bich-Nga Pham[1,2], Thanh Duc Ngo[1,2], Tam V. Nguyen[3], and Vinh-Tiep Nguyen[1,2,*]

[1] *University of Information Technology, Ho Chi Minh City, Vietnam*

[2] *Vietnam National University, Ho Chi Minh City, Vietnam*

[3] *University of Dayton, Dayton, OH 45469, United States*

[*] *Corresponding author: Vinh-Tiep Nguyen* (e-mail: tiepnv@uit.edu.vn),

Abstract—
Instance image segmentation task requires training with abundant annotated data to achieve high accuracy. Recently, conditional image synthesis has demonstrated its effectiveness in generating synthetic data for this task. However, existing image synthesis models face challenges in generating target instances to match the masks with complex shapes. Moreover, others fail to create diverse instances due to utilizing low-context simple text prompts. To address these issues, we propose CAMUL, a framework for context-aware multi-conditional instance synthesis. CAMUL introduces two key innovations: CARP (cross-attention refinement prompting) to enhance the alignment of generated instances with conditional masks, and iCAFF (incremental context-aware feature fusion) to determine the general embeddings of the instances for a more precise context understanding. Our method significantly improves segmentation performance, increasing up to 15.34% *AP on Cityscapes and* 3.34% *AP on the large-scale ADE20K benchmark compared to the baselines.*
Code is available at https://github.com/danhntd/CAMUL.

Scene understanding at the instance level is challenging when requiring models to identify each semantic instance distinguishedly. Such deep learning models normally undergo extensive training on abundant annotated data which is somehow high-cost in instance-wise pixel-level annotations. Recently, the research community addressed the challenge by adopting image synthesis methods embedded into the segmentation framework to leverage the power of generative models to automatically synthesize data for the training scheme. A recent method [1] synthesizes training images and the corresponding annotation masks given text prompts. However, the method focuses on semantic segmentation and cannot straightforwardly apply to address instance segmentation. To resolve this problem, a more recent work [2] proposed a framework to synthesize data in an instance-wise manner based on *multiple conditions of prompts and mask annotation shapes*. Notably, this approach *allows reusing the existing annotation masks* for the training process to save manual effort while leveraging the generation models to increase the diversity of the data at the instance level.

However, the aforementioned approach still faces the following challenges: **First**, the pretrained large multimodal models struggle to accurately generate instances that align with complex mask shapes. It is worth noting that the pixel-wise alignment between the image and its corresponding annotation mask is the

XXXX-XXX © 2024 IEEE
Digital Object Identifier 10.1109/XXX.0000.0000000

key to achieving high accuracy in image segmentation. In current methods, the mask condition is simply taken as one of the inputs to guide the generation process without any specific attention. This causes a misunderstanding of the synthesis model when creating the output instance matching the complex mask shape.
Second, the image synthesis model fails to create diverse instances due to the low context text prompt. The conditional text prompt is constructed based on the class name of the instance located in the simple structural description, i.e., *"a photo of" + <adjective> + <classname>*, resulting in very similar descriptions for different instances. As a consequence, such low-context structural prompts limit the ability of the large pretrained models to generate diverse samples.

In this work, we propose the two corresponding solutions. *To the first problem* of generating the instance matching mask shapes, we propose a cross-attention prompting approach. Unlike previous methods [3, 4], our framework employs the ground-truth mask to control the editing stage instead of the computed attention maps. Our proposal allows such a correct mask to constrain the generation process by adjusting the exact instance regions. As a result, the synthesized instances match the conditional shapes to become a good training sample. *Regarding the second problem* of lacking diversity in generating samples, we propose a novel approach to boost the context comprehension for the generation models. Accordingly, we utilize a feature fusion method, named after iCAFF, to integrate more context information into a prompt. The prompt representation is the fusion of two parts, including the original specific description and the general description of the other neighbor instances. For example, *"a red car"* instance now has more context from the general description of other cars in the category, including *"a black car moving on the road"*, *"a silver car from a front view"*, and *"a white car parking by the road"*, etc. In this way, we bring more contextual information to guarantee the diversity of the generated instances.

To summarize, we propose **CAMUL**, the **C**ontext-**A**ware **Mul**ti-conditional Instance Synthesis framework, designed to address the challenges of limited annotated data in instance segmentation. Our contributions in this work are as follows: **(i)** We introduce CARP as a novel method relying on a cross-attention mechanism to guide the generation process by adjusting the exact instance regions, especially in complicated mask shapes **(ii)** We present iCAFF as a feature fusion method to enrich the information of the conditional prompt to address the limitation in the diversity of synthesized instances. **(iii)** We empirically demonstrate the effectiveness of CAMUL, achieving significant performance improvements in instance segmentation tasks compared to baselines.

The remainder of this paper is organized as follows: Section Related Work reviews relevant works on instance segmentation and conditional image synthesis approaches. The next section details our CAMUL Framework. In Section Experimental Results, we report the results and ablation studies with discussions to prove the effectiveness of our proposals. Finally, Section Conclusion concludes our work.

Related Work

Instance Image Segmentation

Different from semantic segmentation, instance segmentation classifies and clusters each semantic pixel into separate instances with distinct boundaries among instances belonging to the same or different classes. This task addresses problems requiring high-detail information, including scene understanding or supporting other downstream tasks. Such instance-level segmentation methods can be adapted to serve urban scene analysis, i.e., one-stage approach [5, 6], or two-stage approach , typically. Recently, OneFormer [5] is the method with an all-in-one manner that solves multi-tasks of panoptic, semantic, and instance segmentation. Developed on top of Mask2Former [7], FastInst [6] focuses on real-time applications. Instance activation-guided queries, the dual-path update approach, and ground truth mask-guided learning are listed as essential components. In this work, we aim to evaluate the impact of the synthetic data on the instance segmentation models, so we use OneFormer and FastInst as the baselines, as they are commonly utilized recently.

Multi-conditional Image Synthesis

To create images satisfying user intent, multi-conditional image synthesis models rely on multiple sources of input, such as combining text prompts with other conditions like sketches or semantic masks. This approach has been exemplified by generative adversarial networks [8] (GANs) or recently by multimodal diffusion models [3, 4, 9, 10, 11, 12, 13, 14, 15]. For instance, the typical GLIGEN [10] effectively blends textual and visual inputs to guide the image generation process. Such work [11, 12] built on top of the diffusion model also leverages the multimodality to synthesize images. Method [3] provides a solution to control the synthesizing process via a prompt-to-prompt technique to enhance the synthesized result. Recent work [4] addresses the limitation in [3] by proposing a null-text inversion method to serve image editing tasks. By leveraging the complementary

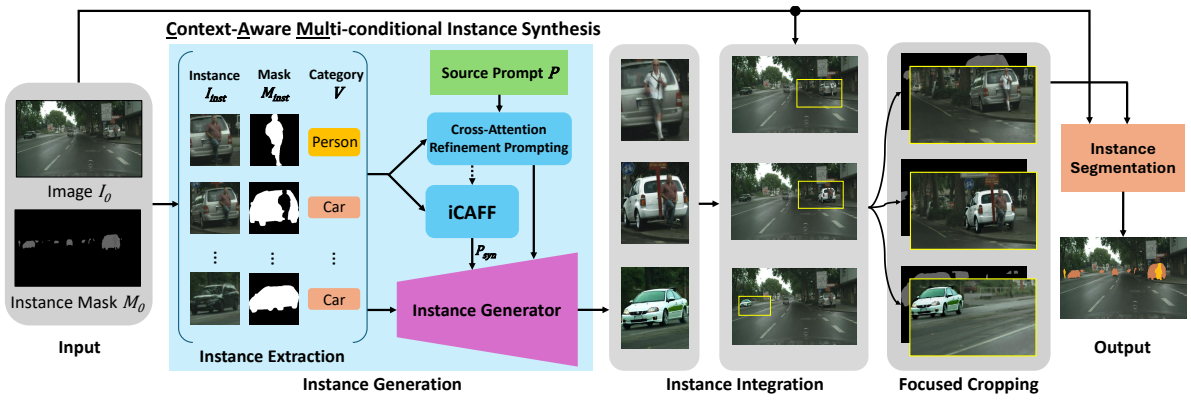

FIGURE 1. Overview of our CAMUL framework. The pipeline allows a pair of image-annotation to be augmented into various variations with category-driven text prompts in terms of boosting the data diversity to serve instance scene understanding.

strengths of different input modalities, such models can generate more contextually appropriate visual content than single-conditional methods. Our work uses this multi-conditional image synthesis approach to address the segmentation task.

Multi-conditional Image Synthesis in Instance Segmentation

Conditional image synthesis methods have been recognized to enrich the diversity and volume of training datasets to boost the segmentation tasks. Recent studies have investigated deep learning-based generation techniques [8, 9]. These approaches maintain the semantic integrity of data while enhancing image diversity through advanced transformations. Methods such as data generation involving GANs [8], and Latent Diffusion [12] have become popular for augmenting datasets with meaningful variations while preserving semantic coherence. Such derivative models of diffusion, such as [3, 4, 10, 11, 12, 13, 14, 15] succeeded in the image generation task with significant performance. In our framework, we utilize [3] with the support of [4] to overcome the existing problem of null-text inversion, which will be discussed in detail. Nonetheless, our framework further provides a mechanism to control the conditional image synthesis process, bringing precise synthesized images to serve segmentation.

CAMUL Framework

The overview of the CAMUL framework is presented in Figure 1 as a context-aware multi-conditional instance synthesis approach in serving the scene understanding task. We propose a pipeline to enrich the input pairs of instance segmentation mask and image at the instance level, given the mask conditions and prompt guidance. Accordingly, a pair of image-instance annotations is taken into the process of instance generation, along with the support of a conditional text prompt. This process provides different variants of augmented instances of the input images. Altogether, the original and its augmented versions are combined via the post-processing of instance integration and focused cropping steps to train the instance segmentation model. We analyze the details of the proposed framework in the following sections.

Instance Generation

Conditional Text Prompt. Recent conditional image synthesis models often misunderstand text prompts and fail to follow them precisely. In this case, our CAMUL utilizes two versions of prompts, including source prompts and target prompts, to manipulate the instance synthesis, as shown in Figure 2-(a). We separate the prompting process into multiple steps to guarantee the understanding of the model. While the source prompts focus on accurately describing the original images, the target prompts specify the desired modifications and guide the editing of the images.

Instead of relying on manually pre-defined text prompts, we leverage a Large Multimodal Model (LMM), which is LLaVA-1.5, to automatically create text prompts from images and formulate this stage as an image captioning task. These generated prompts are referred to as source prompts. LLaVA-1.5 model is capable of understanding images to initialize the text-prompt as a description for the image with a balance between accuracy, inference speed, and computational efficiency. Then, a Large Language Model (LLM), which is LLaMA-3, refines the source prompts by modifying specific attributes mentioned in the prompts,

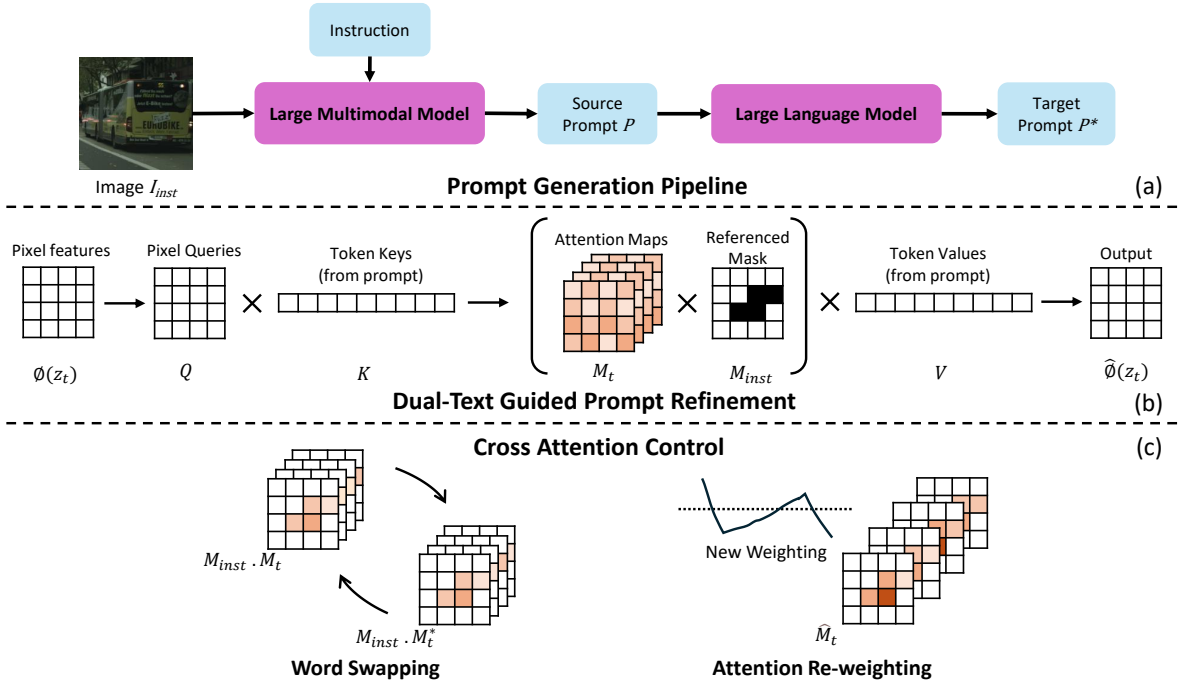

FIGURE 2. Our Cross-Attention Refinement Prompting (CARP). (a) The prompt generation pipeline generates source and target prompts from the input image I_{inst}. (b) Cross-attention layers merge visual and textual embeddings to create spatial attention maps M_t for each text token. A referenced mask M_{inst} taken from annotations then refines these maps to ensure precise editing. (c) In Word Swapping, we replace the source image maps M_t with the target image maps M_{t*}, both guided by a referenced mask M_{inst}. Adjusting the weight of the attention map controls the semantic impact of a word.

such as color and texture, while preserving the overall context. LLaMA-3, with at the time represented state-of-the-art performance, adhered strictly to our instructions and reliably produced the desired target prompts.

In our case, the LLM is utilized to adjust the color of the instance by substituting it with an opposite color to generate the target prompts. This contrasting color may enhance the augmented versions and benefit the visual understanding of the instances. With this approach, we allow the model to focus on the main instance of the image and then augment that instance with conditional prompts. Additionally, considering the context of how the instance appears in the image, we adjust the viewpoint component to enrich the diversity of the augmented image versions. Exploring such LLMs and LMMs, we automate the prompting process to save manual efforts and obtain stable prompting results.

Instance Extraction. In the initial stage, we rely on mask annotations to break down the input image I_0 into instance fragments. This step involves identifying significant instances I_{inst} in the dataset based on the instance annotation mask M_0. V is the set of unique class labels in M_0. We exploit top $K = 3$ instances based on their pixel area ranking as presented in Equation 1. The hyperparameter K is empirically selected in this case to adapt our hardware availability, and such large instances ease the synthesis process of the generative models simultaneously.

$$I_{inst} = Top_K \left\{ \sum_{i=1}^{h} \sum_{j=1}^{w} \delta\left(M_0[i,j], v\right), v \in V \right\}, \quad (1)$$

where:

- h, w are the height and width of the mask M_0
- K is the number of selected instances
- v is a class label in V
- $\delta(x, y)$ is the Kronecker delta function, defined as:

$$\delta(x,y) = \begin{cases} 1, & \text{if } x = y \\ 0, & \text{otherwise} \end{cases}$$

Following the selection, the instances are cropped randomly to remove other irrelevant instances to ensure the generation process concentrates on the selected instance. The corresponding original image I_0

is also cropped in the same manner. These cropped original images and masks are then used as inputs for the subsequent image synthesis component, which aims to seamlessly fill the guided masked areas within the instance images.

Cross-Attention Refinement Prompting (CARP). The extracted instances in the previous stage are the inputs of this CARP module to synthesize instances. Our CARP integrates the strengths of Null-text Inversion [4] and P2P [3] while introducing the impact of the referenced masks to guide the editing stages, as shown in Figure 2-(b). This approach enhances the suitability of the image editing task towards our downstream task of instance segmentation with the intent to reuse the existing annotation masks.

In diffusion-based conditional image editing, cross-attention maps control the interaction between text tokens and image pixels to generate the instance with a given spatial layout. We exploit two key techniques: Word Swapping and Attention Re-Weighting, as described in Figure 2-(c). Word Swapping performs the replacement of attention maps between the original and the target text token, allowing the model to capture directly the changes in prompt guidance without affecting the non-targeted regions. Attention Re-weighting adjusts the impact of the specific tokens by scaling its attention map. Altogether, the two techniques address the key challenges in image editing by enabling fine-grained edits while preserving the integrity of the overall image. This level of control ensures that modifications are both visually coherent and contextually appropriate.

Our solution introduces the ground-truth mask M_{inst}, which is considered as the shape of the instances, to adjust the attention maps precisely. The proposed CARP Algorithm 1 with the integration of the ground-truth mask M_{inst} enables a more tailored fit for this task. In detail, after both attention map M_t and M_t^* at timestep t are produced by the source prompt P and the target prompt P^*, we incorporate a referenced mask M_{inst} into the Edit function along with the aforementioned masks M_{inst}. This setup is essential for executing Word Swapping and Attention Re-weighting, as detailed in Equation 2 and Equation 3. Accordingly, our approach can control a characteristic while preserving the structure of the instance and the overall image integrity.

$$\text{Edit}(M_{inst}, M_t, M_t^*, t) = \begin{cases} M_{inst} \cdot M_t^* & \text{if } t < \tau, \\ M_{inst} \cdot M_t & \text{otherwise,} \end{cases} \quad (2)$$

where τ serves as a timestamp parameter setting the boundary for the injection process. The initial stages of the diffusion process determine the image composition [3]. Therefore, by limiting the injection steps, we can shape the composition of the newly created image while still allowing enough geometric adaptability to align with the new prompt.

Algorithm 1 CARP Workflow

Input: A source prompt P, a target prompt P^*, and a pre-defined mask M_{inst}.
Output: A source image x_{src} and an edited image x_{dst}.
$z_T \sim \mathcal{N}(0, I)$, a unit Gaussian random variable; $DM(z_t^*, P^*, t)$ is the computation of a single step t of the diffusion process, which outputs the noisy image z_{t-1}, and the attention map M_t;
$z_T^* \leftarrow z_T$;
for $t = T, T-1, \ldots, 1$ **do**
  $z_{t-1}, M_t \leftarrow DM(z_t, P, t)$;
  $M_t^* \leftarrow DM(z_t^*, P^*, t)$;
  $\hat{M}_t \leftarrow \text{Edit}(M_{inst}, M_t, M_t^*, t)$;
  $z_{t-1}^* \leftarrow DM(z_t^*, P^*, t)\{M_t \leftarrow \hat{M}_t\}$;
end for
return (z_0, z_0^*)

Let index i correspond a pixel value and j correspond to a text token. We introduce a weighting parameter $\mathcal{E}$ to manipulate the impact of the designated token j^*, $\mathcal{E} \in [-5, 5]$ is empirically selected as the suitable range. The attention maps for all other tokens remain unchanged to maintain the background. The specific modification is as follows:

$$\text{Edit}(M_{inst}, M_t, M_t^*, t)_{i,j} = \begin{cases} \mathcal{E} \cdot M_{inst} \cdot (M_t^*)_{i,j} & \text{if } j = j^*, \\ M_{inst} \cdot (M_t)_{i,j} & \text{otherwise.} \end{cases} \quad (3)$$

As a result, we construct multiple augmented versions of the original image-annotation pairs to serve the downstream instance segmentation task.

Quality Filtering Approach. To resolve the problem of having failed cases in our results, Quality Filtering (QF) is proposed as a post-processing step of our CARP approach. Our goal is to ensure that the final output I_{syn} is close to the desired high-quality of the original image I_{real} to produce high-quality augmented datasets. Accordingly, our **QF**-CARP is designed with a post-processing step of thresholding to decide whether to utilize or eliminate the synthesized samples. We utilize CLIPScore $\mathcal{C}$ [16] and SSIM Score $\mathcal{S}$ [17] as metrics to determine the unqualified generated versions; the larger the value, the better the results. In detail,

CLIPScore eliminates images that are semantically meaningless or distorted with a threshold $\theta_C = 0.8$, i.e., $\mathcal{C}(I_{syn}, I_{real}) < \theta_C$. Simultaneously, SSIM filters images that exhibit structural degradation compared to ground truth with a threshold $\theta_S = 0.75$, i.e., $\mathcal{S}(I_{syn}, I_{real}) < \theta_S$.

Incremental Context-Aware Feature Fusion (iCAFF). As context has a vital impact on image synthesis, we embed more context information into the synthesis process. Our iCAFF is proposed as an advancement to integrate context-aware features incrementally into the prompt generation process. This component enhances the diversity of conditional prompts by incorporating multiple viewpoints taken from its nearest neighbors. By addressing the limitations in the diversity of existing conditional prompt-based image generation models, iCAFF improves the quality of generated instances for image segmentation. Our iCAFF also leverages the generated prompts from CARP. To create comprehensive and contextually diverse prompts, our approach is based on up to n neighbor prompts of the same instance class via an averaging process. In detail, each considered prompt embedding P_{syn} incorporates information from n preceding prompt embeddings of the same class (i.e., $n = 10$) and the prompt embedding for the current instance P_i with the weighted ratios α and β, respectively. The synthesized prompt embedding P_{syn} for the current instance i is calculated as follows:

$$P_{\text{syn}} = \frac{1}{\alpha + \beta}\left[\alpha\left(\frac{\sum_{j=i-n}^{i-1} P_j}{n}\right) + \beta P_i\right], \tag{4}$$

where:

- P_j represents the prompt embedding for the j-th instance.
- P_i is the prompt embedding for the considered instance.
- α, β are the weights assigned to the averaged preceding and current prompt embeddings, respectively. Our experiments select $\alpha = \beta = 1$.

Finally, the computed P_{syn} is used as the conditional text prompt for the instance generation model.

Instance Post-processing

Our instance post-processing includes Instance Integration and Focused Cropping. In **Instance Integration**, we integrate the synthesized instances into the original image I_0 to preserve the context. Then, we propose a **Focused Cropping** component to increase the diversity of the training samples by cropping the synthesized images. We design a focused cropping algorithm allowing the instances to appear at a random position in the images due to a normal distribution $\mathcal{N}$. In detail, we first rely on the instance area S_{ins} to determine the new image S_N as described in Equation 5. From this calculated area, the height h_{target} and width w_{target} of the new image are derived. In Equation 6, we compute the final coordinates Coord(S_N) of the new image. The outcome of this process is the generation of sets of images, each focusing on a selected instance, while maintaining the dataset regulations and standards.

TABLE 1. Statistics on the Number of training instances between the vanilla Cityscapes [18] and our augmented versions. QF-Inst. denotes the Quality Filtering version.

ID	Label	#Original Inst.	#Augmented Inst.	#QF-Inst.
11	Person	17,919	69,932	37,154
12	Rider	1,781	6,950	3,700
13	Car	26,963	118,609	63,271
14	Truck	484	2,210	1,202
15	Bus	379	1,704	857
16	Train	168	789	401
17	Motorcycle	737	2,945	1,587
18	Bicycle	3,675	13,896	7,480
	Total	52,106	217,035	115,652

$$S_N = \lambda S_{\text{ins}}, \lambda \sim \mathcal{N}(\mu, \sigma), \tag{5}$$

$$\text{Coord}(S_N) = \begin{cases} x_{\text{target}} \in [\max(0, x_{\text{ins\_min}} - (w_{\text{target}} - w_{\text{ins}})), x_{\text{ins\_min}}] \\ y_{\text{target}} \in [\max(0, y_{\text{ins\_min}} - (h_{\text{target}} - h_{\text{ins}})), y_{\text{ins\_min}}] \\ h_{\text{target}} = S_N / w_{\text{target}}, w_{\text{target}} = \lambda w_{\text{ins}} \end{cases} \tag{6}$$

where:

- λ follows a normal distribution $\mathcal{N}$ with $\mu = 3$, $\sigma = 1$
- S_{ins} is the area of the instance region

Experimental Results

Settings and Datasets

We selected two state-of-the-art instance segmentation baselines, including OneFormer [5] and FastInst [6] to assess our framework. To test the impact of our proposed method, we use the crop size of the training image with ratio 360×720 instead of $512 \times 1,024$ of the FastInst baseline [6] or 512×512 of the OneFormer baseline [5]). Notably, our models are employed without a CLIP-based backbone to save training memory. Our framework is built on top of Detectron2 [19] framework and follows the original configurations of the published work. In detail, we adopted two GeForce RTX 2080Ti GPUs (11GB per each) and trained with the AdamW optimizer. Our training process occurred with $90K$ iterations with a batch size of 2 and a base

TABLE 2. State-of-the-art comparison on OneFormer [5] with adaptive crop-size evaluated on Cityscapes validation set [18].

Method	Backbone	Synthesis-base	PQ ↑	IoU ↑	AP ↑	AP50 ↑
OneFormer [5]	Mapillary-ConvNext-L	-	48.84	72.58	21.75	40.94
	Swin-L		51.52	74.53	25.68	45.90
CAMUL* (Ours)	Mapillary-ConvNext-L	iCAFF	60.99	78.49	34.08	59.24
		CARP	62.70	80.75	37.09	62.84
		QF-CARP	59.86	79.32	34.18	59.12
	Swin-L	iCAFF	60.04	78.76	35.15	60.51
		CARP	60.20	79.27	34.86	60.40
		QF-CARP	60.41	79.03	35.35	61.10

** denotes our methods based on OneFormer instance segmentation architecture*
All of our reproduced results of OneFormer are w/o CLIP, and w/ a smaller crop size of 360×720
The first, second, and third best results are marked in red, blue, and green, respectively.

TABLE 3. Comparison on FastInst baselines [6] on Cityscapes val-set [18].

Method	Backbone	Synthesis	AP ↑	AP50 ↑
Mask2Former [7]	R50-FPN-D3	-	31.40	55.90
FastInst [6]	R50-FPN-D3	-	35.50	59.00
	R50-FPN-D3*	-	24.93	45.69
	R50-FPN-D3**	-	27.65	49.21
CAMUL (Ours)	FastInst-R50-FPN-D3**	iCAFF	33.73	59.33
		CARP	34.35	60.15
		QF-CARP	35.25	60.43

∗ denotes our results w/o CLIP, and w/ default published settings
∗∗ denotes our results w/o CLIP, and w/ reduced image sizes
The first, second, and third best results are marked in red, blue, and green, respectively.

learning rate of 1×10^{-4}. To synthesize images, we follow the published settings of [3, 4, 11] correspondingly during the generation processes. Table 1 presents the comparison of the number of instances among our proposed methods.

Regarding image generation models, we conducted experiments on models [12] trained on the LAION2b-en dataset and its subset. To train the instance segmentation model, we utilized the Cityscapes [18] and ADE20K [20] datasets with their instance-level annotations as the main training data. Our framework then enriches the diversity and the number of instances found in the training set. Cityscapes [18] consists of $5,000$ images of urban scenes with high-resolution pixel-level annotations of 2048×1024, separated into 19 semantic categories and 8 instance categories. Meanwhile, ADE20K [20] includes over $20K$ images of the general domain with various resolutions. The evaluation process was established on the public validation sets of these benchmarks, including 500 images from Cityscapes and $2,000$ images from ADE20K distributed into all the instance classes.

Evaluation Metrics

To quantitatively evaluate the effectiveness of our proposed framework, we break down the framework and step-by-step evaluate each component. We indirectly rely on the image quality to justify the performance of the generation methods via the recent common metrics, including CLIPScore, FID, SSIM, and PSNR. Notably, better results in these metrics do not guarantee better performance in instance segmentation, but such metrics allow us to obtain the synthesized images with better quality and thus increase the ability to enhance the segmentation accuracy.

To report our instance segmentation performance, we utilize average precision (AP) metrics. Specifically, we report mAP and AP@50 or AP@75. Readers may reach this site[1] for details on the COCO-based evaluation metrics. With experiments conducted based on OneFormer [5], we also report Panoptic Quality (PQ) and Intersection-over-Union (IoU) to evaluate panoptic and semantic segmentation.

State-of-the-art Comparison on Instance Segmentation

Cityscapes Benchmark We reported the established experiments on our CAMUL framework in Table 2 and Table 6. For a fair comparison, we noticed the same backbone architectures, i.e., R50-FPN-D3 for FastInst [6], Mapillary-ConvNext-L, and Swin-L for OneFormer [5]. To this end, our methods improve by large margins compared to the baselines. In detail, our FastInst-based instance segmentation model achieves 33.73%, 34.35%, and 35.25% AP via our iCAFF, CARP, and QF-CARP, respectively. These APs improve over 6.08%, 6.70%, and 7.60% compared to our FastInst reproduced results without CLIP, and with reduced training

[1] https://cocodataset.org/#detection-eval

FIGURE 3. Visualization results on Cityscapes val-set [18] with our FastInst R50-FPN-D3 [6]. The confidence threshold is 0.8. Best viewed with zoomed-in.

TABLE 4. Our performance experimented on FastInst-based model reported on the benchmark of ADE20K val-set [20].

Method	Backbone	Training Data		Detection			Segmentation		
		Original	CARP	AP	AP50	AP75	AP	AP50	AP75
FastInst [6]	R50-FPN-D3*	✓		26.42	39.54	27.54	24.57	39.86	25.17
CAMUL (Ours)	FastInst-R50-FPN-D3*		✓	28.50	41.67	29.81	26.41	41.99	27.09
		✓	✓	**30.16**	**44.04**	**31.46**	**27.91**	**44.26**	**28.60**

∗denotes the results of FastInst w/o CLIP model, and with reduced image size
Our best results are marked in ***bold***

image sizes following OneFormer [5]. To this end, we observe the effectiveness of our QF-CARP method in controlling the generated instance-level data by eliminating failure cases. Regarding the OneFormer-based model, we improve a large margin of AP compared among the baselines, i.e., from 21.75% to approx. 37.09% on ConvNext-L CARP and from 25.68% to approx. 35.35% on Swin-L QF-CARP configurations. However, to adapt to the hardware availability, we reduced the crop size of the input training image down to 360×720, which sacrifices our panoptic and semantic segmentation performance measured on PQ and IoU. However, under the concept of instance segmentation, we achieve comparable AP compared to other methods. Meanwhile, we demonstrate our proposals work well when compared with our baselines and yield promising results when training on full configurations as the other work (i.e. 1025×2049). Exemplary visualization on Cityscapes samples is provided in Figure 3. We provide further investigation on the instance synthesis results in the next section.

ADE20K Benchmark Evaluation

To generally demonstrate the effectiveness of our synthesis method, we established the experiments based on the FastInst [6] and evaluated the method on the large-scale benchmark ADE20K [20]. The results are reported in Table 4. In this experiment, we also validate the impact of the synthetic data and the original real data contributing to the final accuracy of the detection and segmentation task. As reported, with the support of synthetic data from our CARP, the results on the ADE20K validation set yielded over 1.84% on the instance segmentation and 2.08% on the detection results compared to the original FastInst. When combining both synthetic and real data for training, the corresponding mAP increases over 3.34% and 3.74% compared to the baseline. The synthetic data succeeds in supporting the model to understand the real data, which reflects our hypothesis of fulfilling the data distribution via the synthesis approach. In a similar manner to Cityscapes, we apply the instance synthesis CARP to ADE20K, leading to an augmented version of the dataset with $54,800$ images containing $388,048$ instances, expanding around $2.74\times$ the vanilla dataset.

Ablation Instance Synthesis Evaluation

Table 5 demonstrates the progressive enhancement in the quality of synthetic data and the effectiveness of our methods in aligning the synthetic data with the

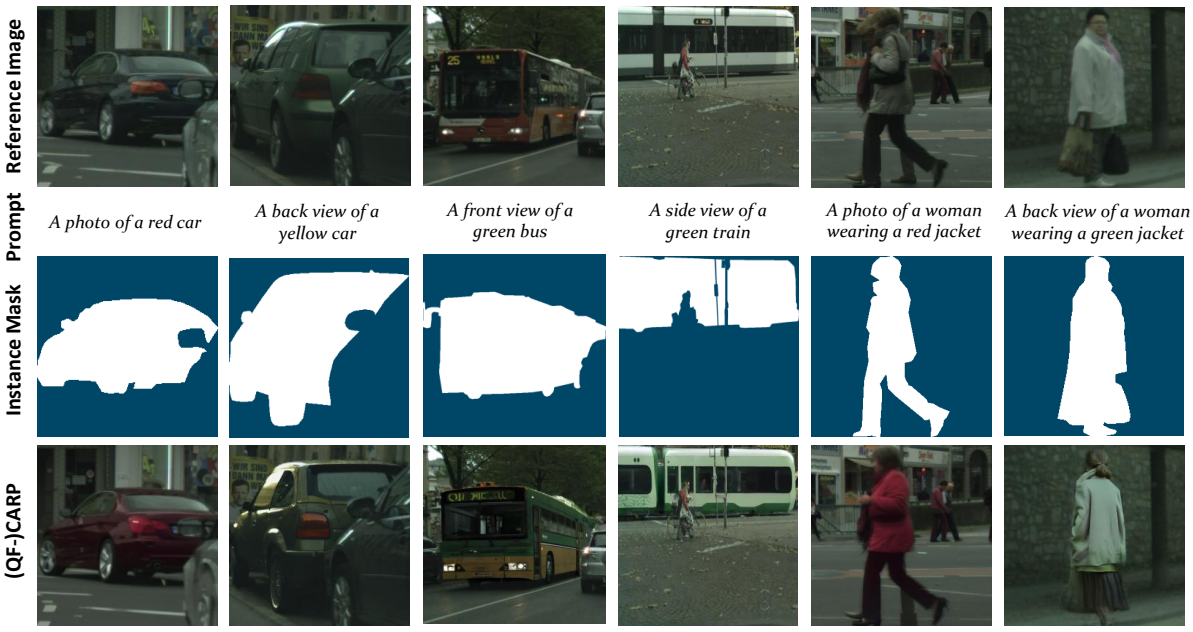

FIGURE 4. Exemplary visualization of the synthesized instances on Cityscapes samples based on our QF-CARP method.

TABLE 5. Comparison of image generators including our iCAFF, CARP, and **QF**-CARP, and other state-of-the-art conditional image generation methods. *The inference speed is measured in second(s).*

Method	Year	CLIPS ↑	FID ↓	SSIM ↑	PSNR ↑	Speed
DiffInpainting [12]	2022	0.81	31.03	0.72	15.95	3.16
BlendedDiff [11]	2022	0.87	16.28	0.90	25.23	11.27
SDEdit [14]	2022	0.58	47.06	0.66	20.87	0.63
InstructPix2Pix [13]	2023	0.83	46.63	0.86	17.44	1.056
GLIGEN [10]	2023	0.79	40.65	0.67	14.39	17.23
ControlNet++ [15]	2024	0.59	133.93	0.22	8.85	2.43
iCAFF (ours)	2025	0.87	22.22	0.89	25.20	11.38
CARP (ours)	2025	0.86	19.88	0.90	27.27	102.60
QF-CARP (ours)	2025	0.90	11.62	0.93	28.24	102.60

The first, second, and third best results are marked in red, blue, and green, respectively.

source data distribution. We indirectly rely on these metrics to judge the quality of the synthetic images, which contribute to the possibility of gaining better downstream segmentation accuracy. The vanilla CARP shows comparable performance in CLIPScore and SSIM compared to BlendedDiff [11], while it surpasses BlendedDiff in PSNR. In detail, BlendedDiff [11] fails to control the synthetic features during the diffusion process, which results in image content that does not reliably fit within the conditional mask, leading to deformations in the edited images. In contrast, the vanilla CARP utilizing DDIM inversion and the modified P2P mechanism can preserve the content of source images while accurately applying changes prompted by the target within the required inpainting region. The **QF**-CARP method demonstrates its effectiveness by outperforming previous methods across all metrics. This indicates that the generated images maintain both low-level and high-level consistency. The visualization of our **QF**-CARP is in Figure 4 with instances extracted from the Cityscapes dataset. Failure may occur when the mask annotations are in complicated shapes.

Discussion. Training a learning model requires a high-quality and comprehensive dataset. Our Quality Filtering approach has demonstrated its superior effectiveness by offering a filter to eliminate the underqualified synthetic cases of CARP. This significant achievement underscores that our approach is capable of syn-

TABLE 6. Inference speed comparison on FastInst baselines [6] on Cityscapes val-set [18].

Method	Backbone	Synthesis	AP ↑	FPS ↑ 2080Ti	FPS ↑ 3090Ti
Mask2Former [7]	R50-FPN-D3	-	31.40	-	-
FastInst [6]	R50-FPN-D3	-	35.50	-	-
	R50-FPN-D3*	-	24.93	5.75	10.94
	R50-FPN-D3**	-	27.65	5.84	11.01
CAMUL (Ours)	FastInst-R50-FPN-D3**	iCAFF	33.73	5.81	10.82
		CARP	34.35	5.83	10.87
		QF-CARP	35.25	5.87	10.89

∗ denotes our results w/o CLIP, and w/ default published settings
∗∗ denotes our results w/o CLIP, and w/ reduced image sizes
The first, second, and third best results are marked in red, blue, and green, respectively.

thesizing an extensive and high-quality dataset. Our assumption in this case is that the better the quality of the training data, the better the model is able to learn from it. Besides, via the cross-attention mechanism in CARP, we ensure the preservation of the source image content while meticulously adhering to the mask during the editing process. This precise alignment between the edited image and the mask not only maintains the integrity of the original content but also enhances the performance of the model in segmentation tasks. Consequently, our (QF-)CARP contributes to more accurate and reliable model training, leading to improved outcomes in downstream instance segmentation tasks.

Conclusion

In this paper, we propose CAMUL - a framework with a context-aware multi-conditional instance synthesis method to address instance-wise scene understanding. The framework is particularly useful in the concept of limited high-cost training annotated data of instance segmentation to enhance the performance of the instance segmentation models. Indeed, we figured out the two main problems of this approach, including the underfitting of the generated instances with mask conditions and the lack of diversity of structural text prompts. To this end, we proposed an incremental prompt fusion method (iCAFF) and adopted the cross-attention refinement prompting (CARP) with a quality filtering component (QF) to respectively get over the phenomenon. We empirically demonstrate the performance of our proposed methods on the Cityscapes benchmark with large margins. Via the instance synthesis, we increase four times the number of instances in the original Cityscapes, resulting in over $200K$ instances serving the training process. Finally, we conduct extensive experiments and ablation studies to prove the effectiveness of our methods over the latest architectures. In the future, we plan to improve the generalization of our proposals by automatically recognizing and embedding the specific features on each instance to allow serving on other data domains.

ACKNOWLEDGMENTS

Thanh-Danh Nguyen was funded by the Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), code VINIF.2024.TS.068.

REFERENCES

1. Q. Nguyen, T. Vu, A. Tran, and K. Nguyen, "Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation," *NeurIPS*, vol. 36, 2023.
2. T.-D. Nguyen, B.-N. Pham, T.-T. D. Vu, V.-T. Nguyen, T. D. Ngo, and T. V. Nguyen, "Instsynth: Instance-wise prompt-guided style masked conditional data synthesis for scene understanding," in *2024 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*. IEEE, 2024, pp. 1–6.
3. A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," *ICLR*, 2023.
4. R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, "Null-text inversion for editing real images using guided diffusion models," in *CVPR*, 2023.
5. J. Jain, J. Li, M. T. Chiu, A. Hassani, N. Orlov, and H. Shi, "Oneformer: One transformer to rule universal image segmentation," in *CVPR*, 2023.
6. J. He, P. Li, Y. Geng, and X. Xie, "Fastinst: A simple query-based model for real-time instance segmentation," in *CVPR*, 2023, pp. 23 663–23 672.
7. B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *CVPR*, 2022, pp. 1290–1299.
8. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *NeurIPS*, vol. 27, 2014.
9. J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, vol. 33, pp. 6840–6851, 2020.
10. Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee, "Gligen: Open-set grounded text-to-image generation," in *CVPR*, 2023.
11. O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in *CVPR*, 2022.

12. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10 684–10 695.
13. T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *CVPR*, 2023, pp. 18 392–18 402.
14. C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "Sdedit: Guided image synthesis and editing with stochastic differential equations," in *ICLR*, 2022.
15. M. Li, T. Yang, H. Kuang, J. Wu, Z. Wang, X. Xiao, and C. Chen, "Controlnet++: Improving conditional controls with efficient consistency feedback," in *ECCV*. Springer, 2024, pp. 129–147.
16. J. Hessel, A. Holtzman, M. Forbes *et al.*, "Clipscore: A reference-free evaluation metric for image captioning," in *EMNLP*, 2021.
17. Z. Wang, A. C. Bovik, H. R. Sheikh *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE TIP*, 2004.
18. M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016, pp. 3213–3223.
19. Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.
20. B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *CVPR*, 2017, pp. 633–641.

Thanh-Danh Nguyen (Graduate Student Member, IEEE) received the B.Sc. (Hons.) and M.Sc. degrees in computer science from the University of Information Technology, VNU-HCM, in 2021 and 2024, respectively, where he is currently pursuing the Ph.D. degree with the Laboratory of Multimedia Communications. His research interests are computer vision and deep learning.

Trong-Tai Dam Vu (Graduate Student Member, IEEE) received his B.Sc. degree in computer science from the University of Information Technology, VNU-HCM, in 2024. He is currently a researcher at the Laboratory of Multimedia Communications at the same institution. His research focuses on deep learning and computer vision, with a particular emphasis on image generation.

Bich-Nga Pham obtained her B.Sc. degree in computer science from the University of Information Technology, VNU-HCM, in 2024. She is pursuing an M.Sc. degree and is working as a researcher at the Laboratory of Multimedia Communications. Her research interests are deep learning and computer vision applications.

Thanh Duc Ngo received a Ph.D. degree from the Graduate University for Advanced Studies (SOKENDAI), in 2013. He has been a Lecturer with the Faculty of Computer Science, University of Information Technology, VNU-HCM, since 2014. His research interests include computer vision and multimedia content analysis.

Tam V. Nguyen (Senior Member, IEEE) received a Ph.D. degree from the National University of Singapore, in 2013. He is an Associate Professor with the Department of Computer Science, University of Dayton. His research interests include artificial intelligence, computer vision, deep learning, multimedia content analysis, and mixed reality. He has authored and co-authored more than 150 papers with over 3,600 citations. His H-index is 31.

Vinh-Tiep Nguyen received the M.Sc. degree from the University of Science, VNU-HCM, in a co-program with the John von Neumann Institute, and the Ph.D. degree from the University of Information Technology, VNU-HCM. He is currently a Lecturer at the University of Information Technology and the Head of the Laboratory of Multimedia Communications (MMLab), VNUHCM. He is passionate about transferring knowledge and research skills to his students. His research interests include computer vision and machine learning.