

CAMUL: Context-Aware Multiconditional Instance Synthesis for Image Segmentation

Thanh-Danh Nguyen¹, Trong-Tai Dam Vu², Bich-Nga Pham³, and Thanh Duc Ngo⁴, *University of Information Technology, Ho Chi Minh City, 72506, Vietnam and Vietnam National University, Ho Chi Minh City, 720325, Vietnam*

Tam V. Nguyen⁵, *University of Dayton, Dayton, OH, 45469, USA*

Vinh-Tiep Nguyen⁶, *University of Information Technology, Ho Chi Minh City, 72506, Vietnam and Vietnam National University, Ho Chi Minh City, 720325, Vietnam*

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 multiconditional 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% average precision (AP) on Cityscapes and 3.34% AP on the large-scale ADE20K benchmark compared to the baselines.

Scene understanding at the instance level is challenging when requiring models to identify each semantic instance in a distinguishing manner. 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 method¹ synthesizes training images and the corresponding annotation masks given

text prompts. However, the method focuses on semantic segmentation and cannot straightforwardly be applied to address instance segmentation. To resolve this problem, a more recent work² 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 (LMMs) 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 key to achieving high accuracy in image segmentation. In current methods, the mask

1070-986X © 2025 IEEE All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.

Digital Object Identifier 10.1109/MMUL.2025.3578972

Date of publication 13 June 2025; date of current version 6 October 2025.

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 (incremental context-aware feature fusion), 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 context-aware multiconditional instance synthesis framework, designed to address the challenges of limited annotated data in instance segmentation. Our contributions in this work are as follows: 1) We introduce CARP (cross-attention refinement prompting) 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. 2) 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. And 3) We empirically demonstrate the

effectiveness of CAMUL, achieving significant performance improvements in instance segmentation tasks compared to baselines.

The remainder of this article is organized as follows: The “Related Work” section reviews relevant works on instance segmentation and conditional image synthesis approaches. The next section details our CAMUL framework. In the “Experimental Results” section we report the results and ablation studies with discussions to prove the effectiveness of our proposals. Finally, the last section 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., a one-stage approach^{5,6} or two-stage approach, typically. Recently, OneFormer⁵ is the method with an all-in-one manner that solves multitasks of panoptic, semantic, and instance segmentation. Developed on top of Mask2Former,⁷ FastInst⁶ 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.

Multiconditional Image Synthesis

To create images satisfying user intent, multiconditional 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 (GANs)⁸ or, recently, by multimodal diffusion models.^{3,4,9,10,11,12,13,14,15} For instance, the typical GLIGEN¹⁰ 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. The method of Hertz et al.³ provides a solution to control the synthesizing process via a prompt-to-prompt technique to enhance the synthesized result. Recent work⁴ addresses the limitation in Hertz et al.³ by proposing a null-text inversion method to serve image editing

tasks. By leveraging the complementary 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.

Multiconditional 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 GANs⁸ and latent diffusion,¹² have become popular for augmenting datasets with meaningful variations while preserving semantic coherence. Such derivative models of diffusion, such as Hertz et al.,³ Mokady et al.,⁴ Li et al.,¹⁰ Avrahami et al.,¹¹ Rombach et al.,¹² Brooks et al.,¹³ Meng et al.,¹⁴ and Li et al.,¹⁵ succeeded in the image generation task with significant performance. In our framework, we utilize Hertz et al.,³ with the support of Mokady et al.,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 multiconditional

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 postprocessing 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 predefined text prompts, we leverage an 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. The LLaVA-1.5 model is capable of

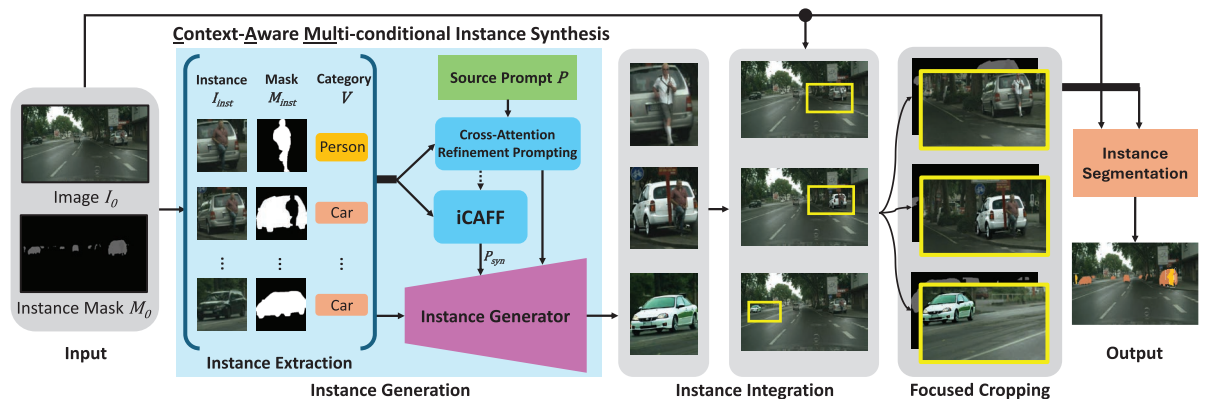


FIGURE 1. Overview of our CAMUL framework. The pipeline allows a pair of image-annotation to be augmented into various variations with category-driven text prompts in terms of boosting the data diversity to serve instance scene understanding.

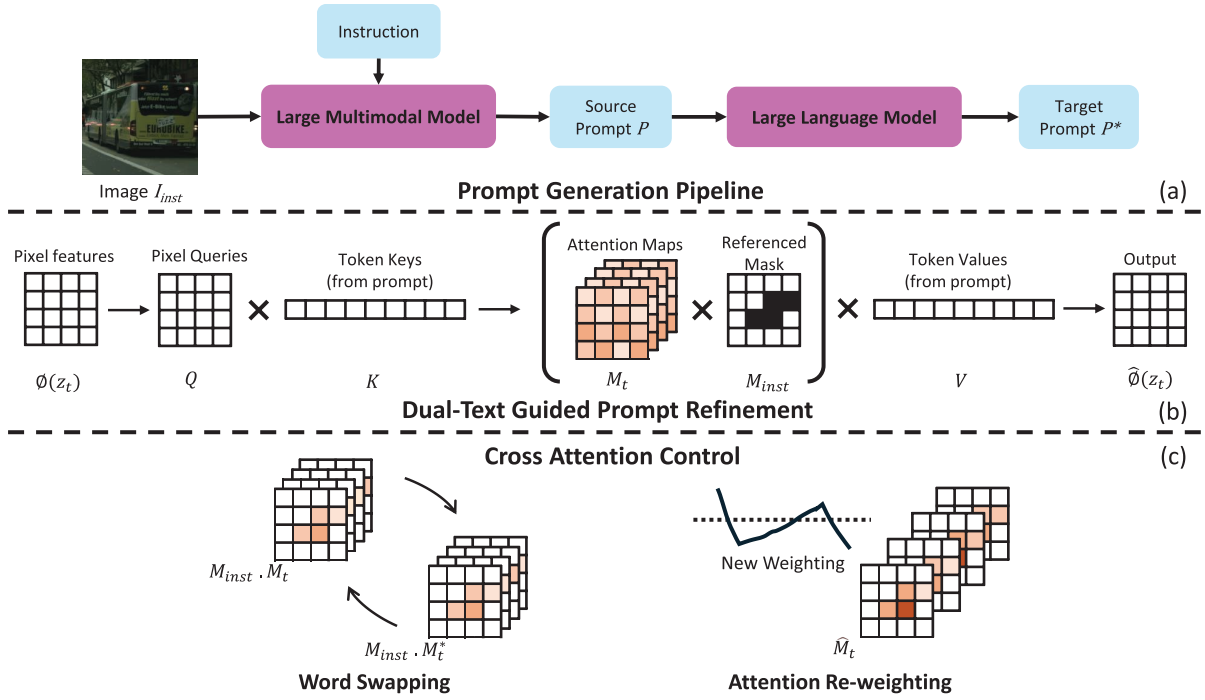


FIGURE 2. Our CARP. (a) The prompt generation pipeline generates source and target prompts from the input image I_{inst} . (b) Cross-attention layers merge visual and textual embeddings to create spatial attention maps M_t for each text token. A referenced mask M_{inst} taken from annotations then refines 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.

understanding images to initialize the text-prompt as a description for the image with a balance between accuracy, inference speed, and computational efficiency. Then, an LLM, which is LLaMA-3, refines the source prompts by modifying specific attributes mentioned in the prompts, 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 (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.

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

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

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 inversion⁴ and peer to peer³ 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 toward 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 reweighting, 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 nontargeted regions. Attention reweighting 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 reweighting, as detailed in (2) and (3). Accordingly, our approach can control a characteristic while preserving

the structure of the instance and the overall image integrity.

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

Algorithm 1: CARP Workflow.

```

1: Input: A source prompt  $P$ , a target prompt  $P^*$ , and a
   predefined mask  $M_{inst}$ 
2: Output: A source image  $x_{src}$  and an edited image
    $x_{dst}$ 
3:  $z_T \sim \mathcal{N}(0, I)$ , 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  $M_t$ ;
4:  $z_T^* \leftarrow z_T$ ;
5: for  $t = T, T-1, \dots, 1$  do
6:    $z_{t-1}, M_t \leftarrow DM(z_t, P, t)$ ;
7:    $M_t^* \leftarrow DM(z_t^*, P^*, t)$ ;
8:    $\hat{M}_t \leftarrow \text{Edit}(M_{inst}, M_t, M_t^*, t)$ ;
9:    $z_{t-1}^* \leftarrow DM(z_t^*, P^*, t) \{M_t \leftarrow \hat{M}_t\}$ ;
10: end for
11: return  $(z_0, z_0^*)$ 
    
```

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.

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:

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

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 postprocessing step of thresholding to decide whether to utilize or eliminate the synthesized samples. We utilize CLIPScore C ¹⁶ and structural similarity index (SSIM) score 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., $C(I_{\text{syn}}, I_{\text{real}}) < \theta_C$. Simultaneously, SSIM filters images that exhibit structural degradation compared to ground truth with a threshold $\theta_S = 0.75$, i.e., $S(I_{\text{syn}}, I_{\text{real}}) < \theta_S$.

Incremental Context-Aware Feature Fusion

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:

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

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 POSTPROCESSING

Our instance postprocessing 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 (5). From this calculated area, the height h_{target} and width w_{target} of the new image are derived. In (6), we compute the final coordinates $\text{Coord}(S_N)$ 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:

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

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

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 OneFormer⁵ and FastInst⁶ to assess our framework. To test the impact of our proposed method, we used the crop size of the training image with ratio 360×720 instead of 512×1024 of the FastInst baseline⁶ 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 Detectron2¹⁹ framework and follows the original configurations of the published work. In detail, we adopted two GeForce RTX 2080Ti GPUs (11 GB per each) and trained with the AdamW optimizer. Our training process occurred with 90,000 iterations with a batch size of 2 and a base learning rate of 1×10^{-4} . To synthesize images, we follow the published settings of Hertz et al.,³ Mokady et al.,⁴ and Avrahami et al.,¹¹ correspondingly during the generation processes. Table 1 presents the comparison of the number of instances among our proposed methods (see also Tables 2 and 3).

TABLE 1. Statistics on the number of training instances between the vanilla Cityscapes¹⁸ and our augmented versions.

ID	Label	No. of Original Instances	No. of Augmented Instances	No. QF-Instnces
11	Person	17,919	69,932	37,154
12	Rider	1781	6950	3700
13	Car	26,963	118, 609	63,271
14	Truck	484	2210	1202
15	Bus	379	1704	857
16	Train	168	789	401
17	Motorcycle	737	2945	1587
18	Bicycle	3675	13,896	7480
Total		52,106	217,035	115,652

QF-Inst: quality filtering version.

TABLE 2. State-of-the-art comparison on OneFormer⁵ 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	45.9
CAMUL* (Ours)	Mapillary-ConvNext-L	iCAFF	60.99	78.49	34.08	59.24
		CARP	62.7	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	60.2	79.27	34.86	60.4
		QF-CARP	60.41	79.03	35.35	61.1

*Denotes our methods based on OneFormer instance segmentation architecture.

All of our reproduced results of OneFormer are without CLIP, and with a smaller crop size of 360×720 . The first, second, and third best results are marked in red, blue, and green, respectively.

Regarding image generation models, we conducted experiments on models¹² trained on the LAION2b-en dataset and its subset. To train the instance segmentation model, we utilized the Cityscapes¹⁸ and ADE20K²⁰ 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. Cityscapes¹⁸ consists of 5000 images of urban scenes with high-resolution pixel-level annotations of 2048×1024 , separated into 19 semantic categories and eight instance categories. Meanwhile, ADE20K²⁰ includes more than 20,000 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 2000 images from ADE20K distributed into all the instance classes.

TABLE 3. Comparison on FastInst baselines⁶ on Cityscapes val-set.¹⁸

Method	Backbone	Synthesis	AP ↑	AP50 ↑
Mask2Former ⁷	R50-FPN-D3	—	31.4	55.9
FastInst ⁶	R50-FPN-D3	—	35.5	59
	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 without CLIP, and with default published settings.

**Denotes our results without CLIP, and with reduced image sizes. The first, second, and third best results are marked in red, blue, and green, respectively.

TABLE 4. Inference speed comparison on FastInst baselines⁶ on Cityscapes val-set.¹⁸

Method	Backbone	Synthesis	AP ↑	FPS ↑	
				2080Ti	3090Ti
Mask2Former ⁷	R50-FPN-D3	-	31.4	—	—
FastInst ⁶	R50-FPN-D3	-	35.5	—	—
	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 without CLIP, and with default published settings.

**Denotes our results without CLIP, and with reduced image sizes.

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

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 peak signal to noise ratio (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 the COCO site^a 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 Tables 2 and 4. 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 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 approximately 37.09% on ConvNext-L CARP and from 25.68% to approximately 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 that 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 FastInst⁶ and evaluated the method on the large-scale benchmark ADE20K.²⁰ The results are reported in Table 5. In this experiment, we also validated 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

^a<https://cocodataset.org/#detection-eval>



FIGURE 3. Visualization results on Cityscapes val-set¹⁸ with our FastInst R50-FPN-D3.⁶ The confidence threshold is 0.8. Best viewed with zoomed-in.

TABLE 5. 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	28.6

*Denotes the results of FastInst without CLIP model, and with reduced image size. Our best results are marked in bold.

TABLE 6. Comparison of image generators including our iCAFF, CARP, and QF-CARP, and other state-of-the-art conditional image generation methods.

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	0.9	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	25.2	11.38
CARP (ours)	2025	0.86	19.88	0.9	27.27	102.6
QF-CARP (ours)	2025	0.90	11.62	0.93	28.24	102.6

The inference speed is measured in second(s). The first, second, and third best results are marked in red, blue, and green, respectively.

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 succeed 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 applied 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× the vanilla dataset.

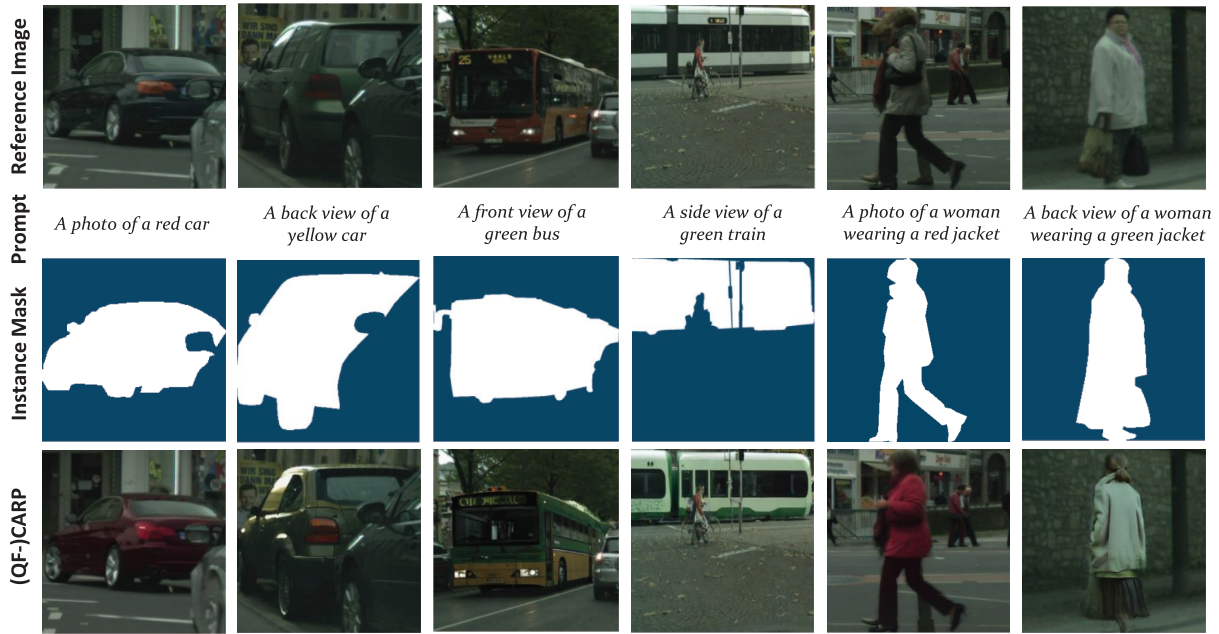


FIGURE 4. Exemplary visualization of the synthesized instances on Cityscapes samples based on our QF-CARP method.

Ablation Instance Synthesis Evaluation

Table 6 demonstrates the progressive enhancement in the quality of synthetic data and the effectiveness of our methods in aligning the synthetic data with the 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 BlendedDiff¹¹ while it surpasses BlendedDiff in PSNR. In detail, BlendedDiff¹¹ 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 a denoising diffusion implicit model inversion and the modified peer to peer 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 QF 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 synthesizing 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 article, we propose CAMUL, a framework with a context-aware multiconditional 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 iCAFF and adopted the CARP with a QF component 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 more than 200,000 instances serving the training process. Finally, we conducted 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, Ph.D. Scholarship Program of the Vingroup Innovation Foundation (VINIF), code VINIF.2024.TS.068. Code is available at <https://github.com/danhntd/CAMUL>.

REFERENCES

1. Q. Nguyen, T. Vu, A. Tran, and K. Nguyen, "Dataset diffusion: Diffusion-based synthetic dataset generation for pixel-level semantic segmentation," in *Proc. 37th Int. Conf. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 76,872–76,892.
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 *Proc. Int. Conf. Multimedia Anal. Pattern Recognit. (MAPR)*, Piscataway, NJ, USA: IEEE Press, 2024, pp. 1–6, doi: [10.1109/MAPR63514.2024.10660775](https://doi.org/10.1109/MAPR63514.2024.10660775).
3. A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross-attention control," in *Proc. 11th Int. Conf. Learn. Representations (ICLR)*, 2023, pp. 1–36.
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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 6038–6047.
5. J. Jain, J. Li, M. T. Chiu, A. Hassani, N. Orlov, and H. Shi, "OneFormer: One transformer to rule universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 2989–2998, doi: [10.1109/CVPR52729.2023.00292](https://doi.org/10.1109/CVPR52729.2023.00292).
6. J. He, P. Li, Y. Geng, and X. Xie, "FastInst: A simple query-based model for real-time instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 23,663–23,672, doi: [10.1109/CVPR52729.2023.02266](https://doi.org/10.1109/CVPR52729.2023.02266).
7. B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 1290–1299, doi: [10.1109/CVPR52688.2022.00135](https://doi.org/10.1109/CVPR52688.2022.00135).
8. I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, 2020, pp. 139–144.
9. J. Ho, A. Jain, and P. Abbeel, "Denosing diffusion probabilistic models," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 6840–6851.
10. Y. Li et al., "GLIGEN: Open-set grounded text-to-image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 22,511–22,521, doi: [10.1109/CVPR52729.2023.02156](https://doi.org/10.1109/CVPR52729.2023.02156).
11. O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 18,187–18,197, doi: [10.1109/CVPR52688.2022.01767](https://doi.org/10.1109/CVPR52688.2022.01767).
12. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 10,684–10,695, doi: [10.1109/CVPR52688.2022.01042](https://doi.org/10.1109/CVPR52688.2022.01042).
13. T. Brooks, A. Holynski, and A. A. Efros, "InstructPix2Pix: Learning to follow image editing instructions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 18,392–18,402, doi: [10.1109/CVPR52729.2023.01764](https://doi.org/10.1109/CVPR52729.2023.01764).
14. C. Meng et al., "SDEdit: Guided image synthesis and editing with stochastic differential equations," in *Proc. 10th Int. Conf. Learn. Representations (ICLR)*, 2022, pp. 1–33.
15. M. Li et al., "ControlNet++: Improving conditional controls with efficient consistency feedback," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer-Verlag, 2024, pp. 129–147.
16. J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "CLIPScore: A reference-free evaluation metric for image captioning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 7514–7528.
17. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004, doi: [10.1109/TIP.2003.819861](https://doi.org/10.1109/TIP.2003.819861).

18. M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 3213–3223, doi: [10.1109/CVPR.2016.350](https://doi.org/10.1109/CVPR.2016.350).
19. Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. "Detectron2 is a platform for object detection, segmentation and other visual recognition tasks." GitHub, 2019. Accessed: Jan. 2, 2025. [Online]. Available: <https://github.com/facebookresearch/detectron2>
20. B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 633–641, doi: [10.1109/CVPR.2017.544](https://doi.org/10.1109/CVPR.2017.544).

THANH-DANH NGUYEN is pursuing a Ph.D. degree with the Laboratory of Multimedia Communications at the University of Information Technology, Ho Chi Minh City, 72506, Vietnam. His research interests include computer vision and deep learning. Nguyen received his M.Sc. degree in computer science from the University of Information Technology. He is a Graduate Student Member of IEEE. Contact him at danhnt.ncs012024@grad.uit.edu.vn.

TRONG-TAI DAM VU is a researcher in the Laboratory of Multimedia Communications at the University of Information Technology, Ho Chi Minh City, 72506, Vietnam. His research interests include deep learning and computer vision, with a particular emphasis on image generation. Vu received his B.Sc. degree in computer science from the University of Information Technology. He is a Graduate Student Member of IEEE. Contact him at taidvt@uit.edu.vn.

BICH-NGA PHAM is pursuing an M.Sc. degree and is working as a researcher in the Laboratory of Multimedia Communications at the University of Information Technology, Ho Chi Minh City, 72506, Vietnam. Her research interests include deep learning and computer vision applications. Pham received her B.Sc. degree in computer science from the University of Information Technology. Contact her at ngapb@uit.edu.vn.

THANH DUC NGO is a lecturer with the Faculty of Computer Science, University of Information Technology, Ho Chi Minh City, 72506, Vietnam. His research interests include computer vision and multimedia content analysis. Ngo received his Ph.D. degree from the Graduate University for Advanced Studies. Contact him at thanhnd@uit.edu.vn.

TAM V. NGUYEN is an associate professor with the Department of Computer Science, University of Dayton, Dayton, OH, 45469, USA. His research interests include artificial intelligence, computer vision, deep learning, multimedia content analysis, and mixed reality. Nguyen received his Ph.D. degree from the National University of Singapore. He is a Senior Member of IEEE. Contact him at tamnguyen@udayton.edu.

VINH-TIEP NGUYEN is a lecturer at the University of Information Technology and the Head of the Laboratory of Multimedia Communications at the University of Information Technology, Ho Chi Minh City, 72506, Vietnam. His research interests include computer vision and machine learning. Nguyen received his Ph.D. degree from the University of Information Technology. He is the corresponding author of this article. Contact him at tiepvn@uit.edu.vn.