Task-Oriented Data Synthesis and Control-Rectify Sampling for Remote Sensing Semantic Segmentation

Supplementary Material

1. Differences Between CRFM and Inversion-Free Editing

Alg. 1 shows the pseudo-code of Control-Rectify Flow Matching (CRFM), which is designed to modulate the trajectory during the early RF process, guiding it toward the target conditional distribution and thereby improving the semantic alignment of the final generated images.

The concept of steering the rectified flow (RF) trajectory has also been similarly applied in image editing tasks. FlowEdit [4] modulates the trajectory of the RF by calculating the vector field difference between the source and target images at each time step. This approach obviates the need for latent inversion, and directly establishes the transition trajectory between the source and target images for image editing tasks. SplitFlow [10] further introduces a flow decomposition and aggregation framework, designed to address the inherent issues of gradient entanglement and semantic conflict within complex target prompts, which mitigates interference between attributes while maintaining overall semantic consistency.

The trajectories established by the aforementioned two methods represent a transformation process from the source data distribution $\boldsymbol{p}_{\text{data}}^{\text{src}}$ to the target data distribution $\boldsymbol{p}_{\text{data}}^{\text{tgt}}$. This process can be formally expressed in the form of the following ODE:

$$dz^{edit} = v_t^{\Delta}(z_t^{src}, z_t^{tgt})dt, \tag{1}$$

where z^{edit} denotes the latent state on the editing trajectory, while z_t^{src} and z_t^{tgt} represent the interpolated latent representations of the source and target states, respectively, at time $t \in [0, 1]$. Specifically, the velocity field difference $v_t^{\Delta}(z_t^{src}, z_t^{tgt})$ is defined as the subtraction of the source velocity field from the target velocity field, where $v(z_t, t)$ is the predicted velocity at latent state z_t and time t:

$$v_t^{\Delta}(z_t^{src}, z_t^{tgt}) = v(z_t^{tgt}, t) - v(z_t^{src}, t). \tag{2}$$

In contrast, our CRFM modulates the trajectory that connects the noise distribution $z_1 \sim \mathcal{N}(0, 1)$ to the conditional data distribution $z_0 \sim p(z_{data}|C^m)$ which is controlled by a semantic mask C^m.The probability path of this process can be constructed by solving the following ODE:

$$dz^{ctrl} = v_t^{CRFM}(z_t, t, C^m)dt, \tag{3}$$

where dz^{ctrl} is the latent state on the conditional trajectory. The term $v_t^{CRFM}(z_t, t, C^m)$ is the vector field adjusted by

Algorithm 1: Control-Rectify Flow Matching

Input : Vector field predictions v_θ, Timestep latents z_t, Noise schedule σ_t, VAE Decoder $\mathcal{D}$, Conditional model Φ, Ground truth labels Y_{gt}

Output: Rectified vector field v^*

$z_0' \leftarrow z_t - \sigma_t \cdot v_\theta$;

$x_0' \leftarrow \mathcal{D}(z_0')$

$\hat{Y} \leftarrow \Phi(x_0')$;

$\mathcal{L}_{cond} \leftarrow \text{CrossEntropy}(\hat{Y}, Y_{gt})$;

$g \leftarrow \nabla_{v_\theta} \mathcal{L}_{cond}$;

$v^* \leftarrow v_\theta - g$;

return v^*

CRFM based on the semantic mask condition C^m at the current latent state z_t (at time t). CRFM vector field can be decomposed into two distinct components:

$$v_t^{CRFM}(z_t, t, C^m) = v_t(z_t, t, C^m) + v_t^{rec}(z_t, t, C^m). \tag{4}$$

Here, $v_t(z_t, t, C^m)$ represents the output vector field of RF model, conditioned on the semantic mask C^m and computed at the current latent state z_t and time t. $v_t^{rec}(z_t, t, C^m)$, is an estimated vector derived from the gradient, calculated using a Segmentation Network $\mathcal{S}$ and the semantic mask $\boldsymbol{C^m}$ on the pre-synthesized image z_0^t of z_t. This component is formally defined as:

$$v_t^{rec}(z_t, t, C^m) \approx \alpha \nabla_v \mathcal{L}_{CE}(\mathcal{S}(z_0^t), C^m), \tag{5}$$

$$z_0^t = z_t - \sigma_t v_t(z_t, t, C^m), \tag{6}$$

where α is a manually set hyperparameter, σ_t is the schedule coefficient that governs the blend ratio between the noise and data during the forward process of RF. Furthermore, $\mathcal{L}_{CE}(\mathcal{S}(z_0^t), C^m)$ calculates the semantic loss between the output of the Segmentation Network $\mathcal{S}(z_0^t)$ and the semantic mask C^m.

Beyond these technical differences in trajectory construction, the core motivations of the two approaches are fundamentally distinct. FlowEdit and related editing methods are built on the principle that an effective editing algorithm should minimally alter the source image while faithfully transporting it toward the target distribution. Their objective is to construct a minimal and direct transition between p_{data}^{src} to p_{data}^{tgt}.

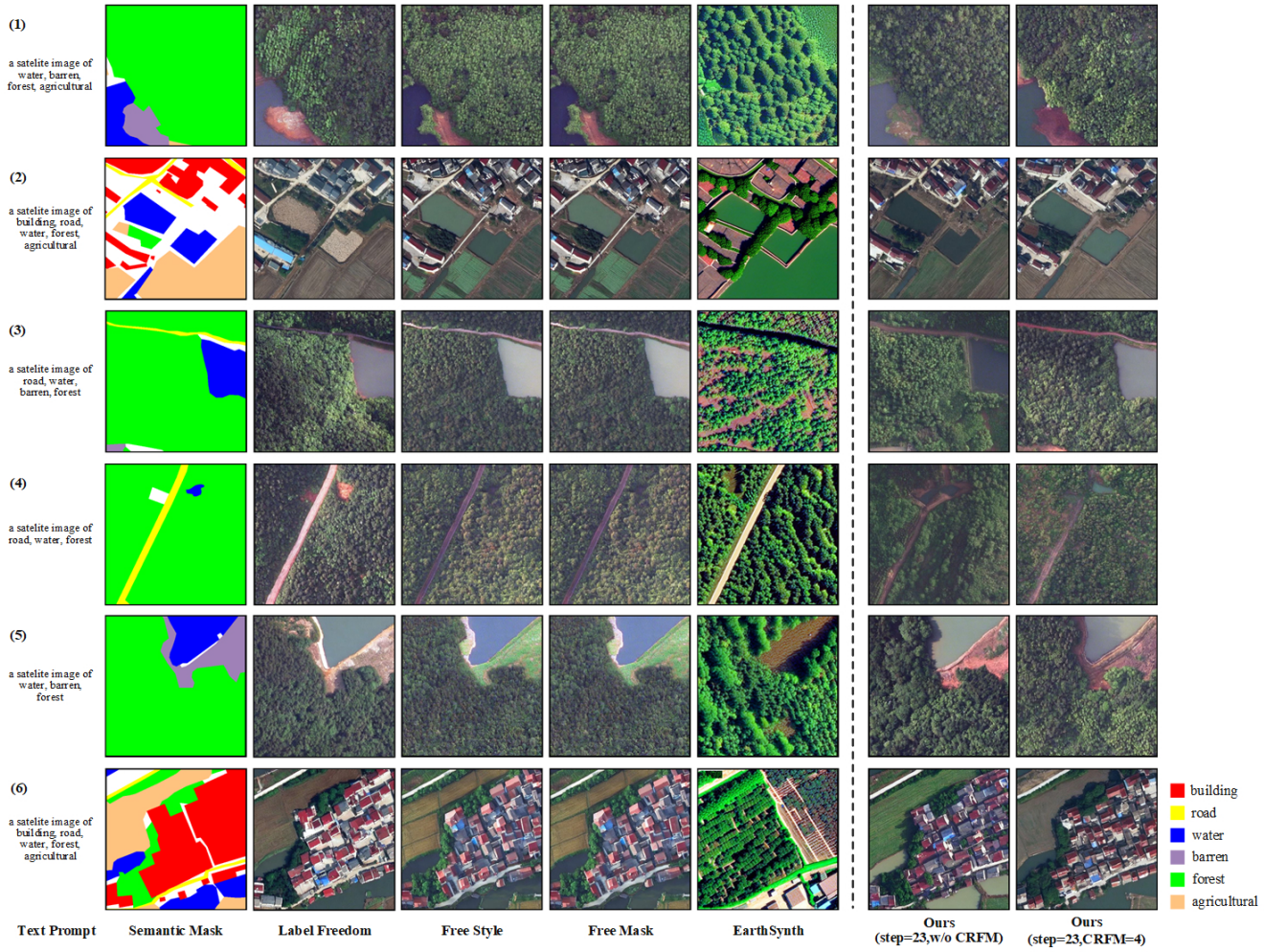

Figure 1. Qualitative comparison for mask-to-image generation by various methods.

In sharp contrast, CRFM aims to modulate the probability trajectory during the RF process for any initial state $z_1 \sim \mathcal{N}(0, 1)$, ensuring that the resulting generated state z_0 more closely aligns with the target data distribution $p(z_{data}|C^m)$.

2. Additional Experiment

2.1. Visualization of Comparative Results

We provide the visualization results of different methods on the LoveDA [7] dataset, including LabelFreedom [13], Freestyle [8], FreeMask [9], and EarthSynth [5], all of which are based on ControlNet [12]. As shown in Figure 1, the ControlNet architecture achieves stronger edge control because the feature outputs from its semantic mask control flow are directly added into the intermediate layers of the image diffusion branch via skip connections. This localized injection mechanism provides tight spatial conditioning.

In contrast, the MMDiT architecture processes the concatenated features of text and the semantic mask through a global Transformer-based attention mechanism. This architectural difference enables MMDiT to model long-range dependencies and better capture global context, thereby resulting in more natural and aesthetically pleasing generation and higher semantic consistency in complex scenes. By optimizing the vector field's direction in the pixel-level latent space, CRFM significantly enhances the precision of edge control, thereby complementing the shortcomings of MMDiT in this regard.

2.2. Sensitivity Analysis of CRFM on Pre-Trained Semantic Segmentation Models

For the semantic segmentation model, we employ PSPNet [14]. This choice is supported by the comparative analysis in TISynth [2], which evaluates architectures including PSPNet, Segmenter [6], and Mask2Former [1]. Their findings indicates that CNN-based architectures demonstrate a superior performance advantage over transformer-based ar-

Table 1. The performance of downstream tasks with varying pre-trained semantic segmentation models on CRFM.

Method	Seg. Model-1			Seg. Model-2			Seg. Model-3		
	OA	mIoU	mAcc	OA	mIoU	mAcc	OA	mIoU	mAcc
Pre-trained Seg. model (baseline)	74.27	45.27	56.44	69.11	40.76	52.98	67.98	30.12	40.48
Downstream task(step=18, CRFM=2)	75.86	47.99	59.46	75.11	47.30	59.89	75.07	46.35	58.64
Downstream task(step=18, CRFM=4)	76.24	48.93	60.54	75.46	48.35	60.82	75.47	47.65	60.19
Downstream task(step=18, CRFM=6)	**76.02**	**49.57**	**62.19**	**75.66**	**48.92**	**61.14**	**75.58**	**48.44**	**60.67**

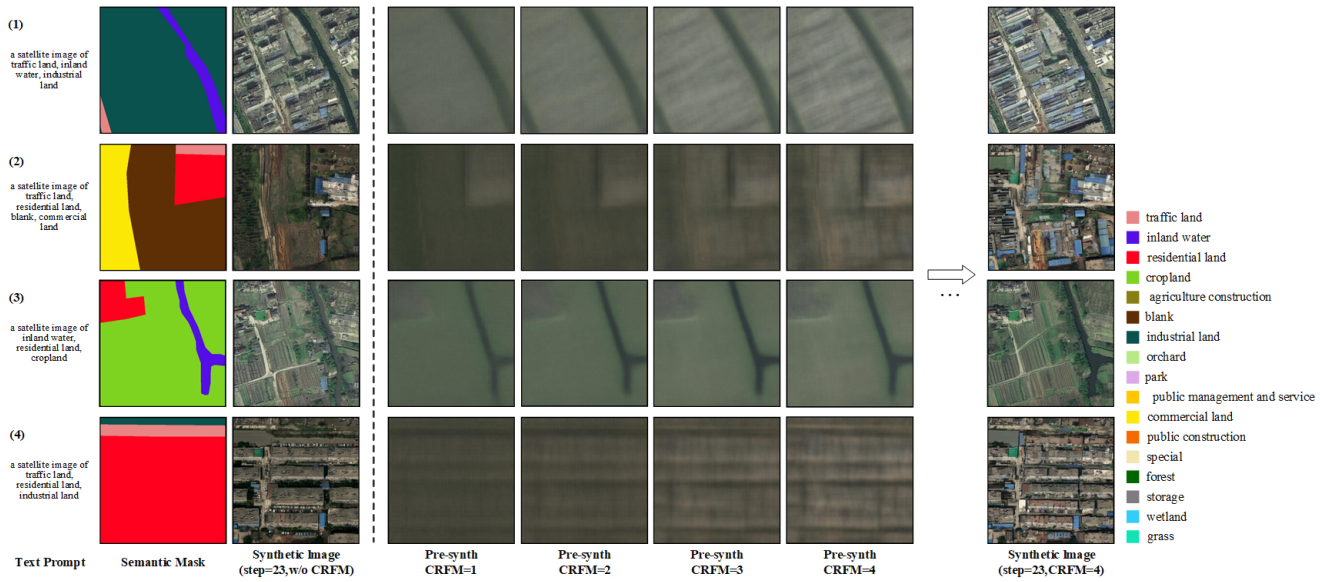

Figure 2. Visualization results of pre-synth images with different CRFM steps on the FUSU.

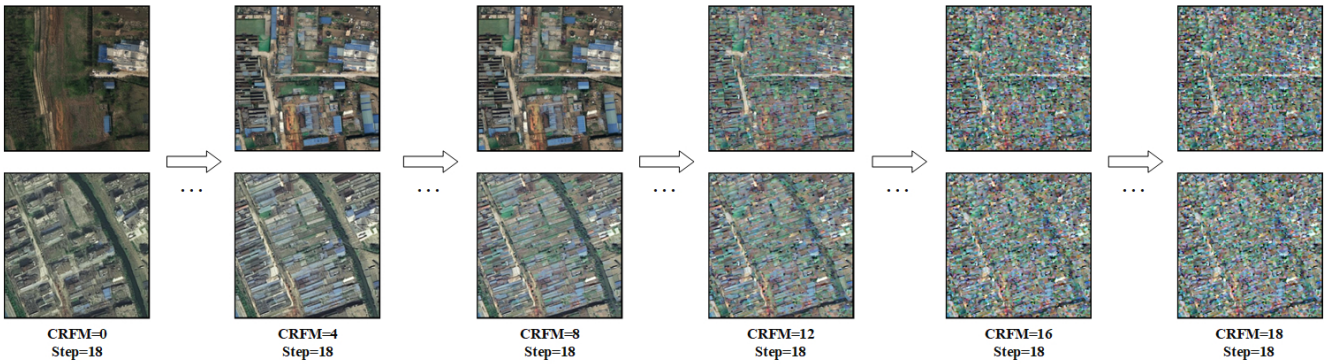

Figure 3. Visual comparison of generated samples with varying CRFM steps.

chitectures for small datasets like FUSU-4k [11].

Because CRFM relies on the semantic segmentation model to compute the semantic loss between its prediction and the ground truth mask, the quality of the pre-trained model may directly affects CRFM performance. To varify the sensitivity of CRFM on pre-trained semantic segmentation models, we design three segmentation models with different accuracies: Seg. Model-1 is trained on the full FUSU-4k dataset. Seg. Model-2 is trained on a 2k subset and Seg. Model-3 is trained on a 1k subset.

As shown in Table 1, we first report the performance of the pre-trained semantic segmentation models. Using these models as the evaluation modules within CRFM, we compare CRFM under different number of adjustment steps

Table 2. The improvement of Control-Rectify Flow Matching on rare categories.

Method	Orchard		Pub. Mngmt.		Storage	
	IoU	Acc	IoU	Acc	IoU	Acc
Baseline [14]	30.01	40.65	39.36	48.60	32.72	39.33
FreeMask [9]	37.71	42.90	41.62	54.50	32.60	43.95
SD v3.5 [3]	37.73	45.26	41.58	54.45	33.86	49.54
Ours	**41.88**	**55.37**	**42.79**	**58.00**	**40.10**	**56.91**

(i.e.,2, 4 and 6), while fixing the sampling step at 18. When utilizing lower-performing pre-trained models, the effectiveness of CRFM is slightly reduced. However, as the number of adjustment steps increases, the final performance consistently surpasses the baseline. This indicates that although the pre-trained semantic segmentation models vary in accuracy, they all possess a certain level of semantic discrimination capability. Such information is sufficiently reliable to guide and correct the sampling trajectory during the early sampling stage. Therefore, although CRFM is somewhat sensitive to the performance of the pre-trained segmentation model, this influence is limited because CRFM operates only in the early sampling stage. Consequently, it consistently delivers improvements over the baseline.

2.3. Performance Gains on Rare-Classes

To evaluate the robustness of CRFM, we analyse experiments on three rare classes: Orchard, Public Management (Pub. Mngmt.), and Storage. As shown in Table 2, our method achieves substantial performance leaps, consistently outperforming state-of-the-art baselines like SD v3.5 [3] and FreeMask [9]. Notably, in the Storage domain, CRFM improves IoU from 33.86% to 40.10% and significantly boosts Accuracy (Acc) from 49.54% to 56.91%.

These gains provide strong empirical evidence for our CRFM mechanism. In data-scarce scenarios, the vanilla vector field v_θ or noise ϵ_θ often provides biased or blurry guidance due to the scarcity of samples. By introducing the rectification term $g = \nabla_{v_\theta}\mathcal{L}_{cond}$, CRFM actively steers the flow trajectories v^* toward the precise semantic manifold. The marked improvement in Acc—reaching 58.00% in Pub. Mngmt.—underscores that our CRFM effectively restores category boundaries and high-frequency details that are typically lost in standard flow-matching trajectories and noise-based diffusion processes.

2.4. Analysis of Sampling with CRFM

To better understand the effect of CRFM on the flow trajectory, we visualize the pre-synth images generated by the CRFM algorithm during sampling, with the total sampling steps set to 18 and the number of CRFM adjustments set to 4. As shown in Figure 2, the generated pre-synth images increasingly align with the semantic structure of the ground-truth masks as the CRFM adjustments proceed. This indicates that CRFM is capable of adjusting the flow trajectory of the noise toward $p(z_{data}|C^m)$, a conclusion further supported by the comparison between generated images with and without CRFM.

However, we observe that setting the number of CRFM adjustments excessively high—typically exceeding 50% of the total sampling steps leads to mode collapse. This phenomenon occurs because the generated images over-optimize to satisfy the pre-trained segmentation model's perception, thereby resulting in a noticeable degradation of human-perceived visual quality. Figure 3 presents the visualization results for a total of 18 sampling steps, while varying the CRFM adjustment count from 0 to 18. A slight mode collapse appears at CRFM = 8, becomes severe at CRFM = 12, and when the CRFM frequency approaches the full sampling length, the outputs degenerate into adversarial-like samples that completely lose visual quality.

References

[1] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 2

[2] Runmin Dong, Shuai Yuan, Litong Feng, Jinxiao Zhang, Weijia Li, Mengxuan Chen, Bin Luo, Wayne Zhang, and Haohuan Fu. Transferable image synthesis for remote sensing semantic segmentation via joint reference-semantic fusion. *Information Fusion*, 127:103839, 2026. 2

[3] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 4

[4] Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19721–19730, 2025. 1

[5] Jiancheng Pan, Shiye Lei, Yuqian Fu, Jiahao Li, Yanxing Liu, Yuze Sun, Xiao He, Long Peng, Xiaomeng Huang, and Bo Zhao. Earthsynth: Generating informa-

tive earth observation with diffusion models. *arXiv preprint arXiv:2505.12108*, 2025. 2

[6] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021. 2

[7] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021. 2

[8] Han Xue, Zhiwu Huang, Qianru Sun, Li Song, and Wenjun Zhang. Freestyle layout-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14256–14266, 2023. 2

[9] Lihe Yang, Xiaogang Xu, Bingyi Kang, Yinghuan Shi, and Hengshuang Zhao. Freemask: Synthetic images with dense annotations make stronger segmentation models. *Advances in Neural Information Processing Systems*, 36:18659–18675, 2023. 2, 4

[10] Sung-Hoon Yoon, Minghan Li, Gaspard Beaudouin, Congcong Wen, Muhammad Rafay Azhar, and Mengyu Wang. Splitflow: Flow decomposition for inversion-free text-to-image editing. *arXiv preprint arXiv:2510.25970*, 2025. 1

[11] Shuai Yuan, Guancong Lin, Lixian Zhang, Runmin Dong, Jinxiao Zhang, Shuang Chen, Juepeng Zheng, Jie Wang, and Haohuan Fu. Fusu: A multi-temporal-source land use change segmentation dataset for fine-grained urban semantic understanding. *Advances in Neural Information Processing Systems*, 37:132417–132439, 2024. 3

[12] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2

[13] Chenbo Zhao, Yoshiki Ogawa, Shenglong Chen, Zhehui Yang, and Yoshihide Sekimoto. Label freedom: Stable diffusion for remote sensing image semantic segmentation data generation. In *2023 IEEE International Conference on Big Data (BigData)*, pages 1022–1030. IEEE, 2023. 2

[14] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2, 4