DuDoST: A Dual-Domain Self-Denoising Transformer for Robust Semantic Segmentation

Jinming Li[1], Wenzhu Yang[1, 2, *], Shengbo Wang[1]
[1]School of Cyber Security and Computer, Hebei University, Baoding, China
[2]Machine Vision Engineering Research Center, Hebei University, Baoding, China
[*]Corresponding Author. Email: wenzhuyang@hbu.edu.cn

***Abstract*—Noise interference and semantic instability are key factors that limit the generalization ability and prediction consistency of Transformer-based semantic segmentation models. Although the global modeling capability of Transformers helps alleviate local ambiguity, unconstrained global attention often introduces noise during the encoding stage and amplifies semantic instability. In addition, the lack of explicit modeling of predicted mask structures causes errors to continuously accumulate during decoding and upsampling, ultimately affecting the overall consistency of segmentation results. To address these issues, we propose a Dual-Domain Self-Denoising Transformer (DuDoST), which effectively suppresses background noise and enhances foreground region representations while exhibiting stronger structural awareness and self-repair capability. The Hierarchical Semantic Prior Gating (HSPG) embedded in DuDoST effectively alleviates the interference of background noise in multi-scale features on the Transformer decoding process and enhances the model's attention to semantically relevant regions. The Mask-domain Diffusion Denoiser (MaDD) introduces conditional diffusion modeling in the mask domain, significantly improving the structural consistency and robustness of segmentation results without requiring additional post-processing. Extensive experiments on the Cityscapes and ADE20K datasets demonstrate that the proposed method achieves state-of-the-art performance, obtaining 82.9% mIoU on Cityscapes and 50.5% mIoU on ADE20K.**

***Index Terms*—Transformer, Diffusion Modeling, Noise-aware Learning, Dual-domain Denoising**

I. Introduction

Semantic segmentation essentially aims to classify each pixel in an image into a predefined semantic category, enabling machines to recognize and parse visual scenes in a manner similar to humans. In recent years, Transformer-based segmentation methods have become a mainstream approach in semantic segmentation. Compared with traditional CNN-based networks, Transformer-based segmentation methods rely on global modeling, making them more sensitive to the quality of input features. As Transformers have been increasingly applied to dense prediction tasks, related issues have also emerged: while global attention and discrete token representations enhance contextual modeling capability, they also amplify the interference of background noise and semantic instability. This limitation becomes more pronounced in complex scenarios.

Early Transformer-based segmentation methods, such as SETR [1], DPT [2], Segmenter [3], SegFormer [4], MaskFormer [5], and K-Net [6], establish semantic consistency at the token level through global self-attention. Although this strategy alleviates local ambiguity, issues such as boundary noise and semantic drift become more pronounced under complex conditions. Subsequently, a series of methods attempted to design multi-scale feature reorganization and hierarchical encoders. however, their suppression of noise is mostly achieved in an indirect manner.

With further research, instability itself has been regarded as a core issue that should be explicitly modeled and constrained. Methods such as Mask2Former [7] and DAFormer [8] alleviate semantic fluctuations by improving attention mechanisms and introducing iterative error correction. MP-Former [9], OneFormer [10], and ProDA [11] point out that factors such as cross-layer prediction drift, low query utilization, and conflicts among classification objectives are important causes of unstable Transformer-based segmentation performance. Inspired by these findings, in recent years, generative modeling and diffusion-based methods such as UniGS [12] have explicitly modeled the structural rationality of output distributions through a progressive denoising process, providing stronger prior constraints for suppressing noise and enhancing semantic consistency. This suggests that shifting the focus of segmentation stability from training heuristics to modeling and prior design may become a new research direction.

In summary, relying solely on global attention or structured mask prediction is insufficient to fundamentally resolve these issues. Existing methods still lack a unified solution that can actively suppress irrelevant semantic interference during the feature modeling stage and further repair structural defects during the mask prediction stage. To this end, we propose a Dual-Domain Self-Denoising Transformer (DuDoST), which addresses noise interference and semantic instability in Transformer-based segmentation in a complementary and systematic manner from both the feature space and the mask space, and its effectiveness is ultimately validated on the Cityscapes and ADE20K datasets.

The main contributions of this paper are summarized as follows:

- We propose a Dual-Domain Self-Denoising Transformer (DuDoST), which addresses noise interference and semantic instability in semantic segmentation by jointly optimizing from two complementary perspectives, namely the feature space and the mask space.
- We propose a Hierarchical Semantic Prior Gating (HSPG), which explicitly incorporates semantic priors

in the feature space to guide the model to focus on semantically relevant regions across multi-scale features, thereby effectively suppressing background noise and enhancing foreground representations.

- We propose a Mask-domain Diffusion Denoiser (MaDD), which introduces diffusion modeling in the mask space and explicitly incorporates the structural recovery process into the training objective, enhancing the structural consistency and robustness of segmentation results.
- Extensive experiments on the Cityscapes and ADE20K datasets demonstrate that the proposed method outperforms existing complex state-of-the-art approaches.

II. Related Work

A. Semantic Guidance in Feature Space

To alleviate background interference and semantic confusion in semantic segmentation under complex scenarios, a large body of research enhances the representation of semantically relevant regions by modulating feature responses or guiding contextual aggregation. PSPNet [13], the DeepLab [14] series, as well as EncNet [15] and GCNet [16] improve semantic consistency by modeling global or scene-level semantic context. OCNet [17] and OCRNet [18] enhance segmentation performance in high-resolution scenes by explicitly constructing object-level or region-level semantic representations. CCNet [19], ANLNet [20], and DANet [21] strengthen the consistency of distant regions belonging to the same category by explicitly modeling long-range dependencies or global relationships. PCANet [22] introduces part-level class activation cues to construct attention constraints, achieving improved segmentation performance in scenarios with category confusion. In summary, existing studies are mainly limited to semantic prior modeling, and constructing stable, explicit, and multi-scale consistent semantic priors in the feature space remains an important research direction.

B. Structural Modeling and Refinement in Mask Space

To address the difficulty of adequately modeling object geometry and semantic consistency in semantic segmentation under complex scenarios, a large body of research has gradually explored two complementary perspectives, namely structural modeling of mask prediction and mask refinement. Mask R-CNN [23], MaskFormer [5], and Mask2Former [7] introduce object-level mask branches, set prediction, and masked attention mechanisms to enable a transition of mask prediction from pixel-wise classification to object- and set-level structured mask modeling. Various traditional and non-diffusion-based mask refinement methods, such as DenseCRF [24], RefineNet [25], PointRend [26], SegFix [27], and BPR [28], have proven effective in improving boundary alignment and fine-grained details of predicted masks. These observations indicate that combining strongly structured mask prediction models with mask-domain diffusion-based refinement mechanisms has significant research value at present.

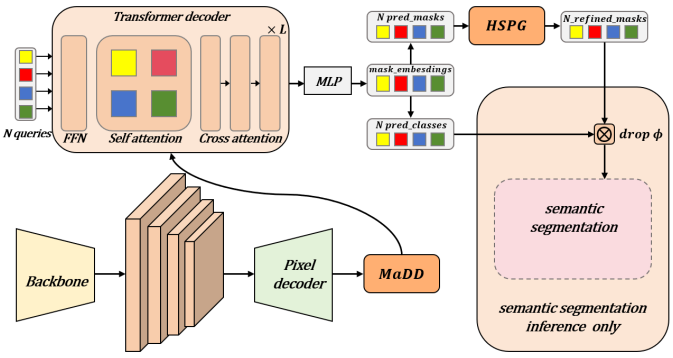

Fig. 1. Overall architecture of the proposed DuDoST framework.

III. Method

A. Overview

To alleviate the poor generalization ability and prediction consistency of Transformer-based semantic segmentation models caused by noise interference and semantic instability, we propose a Dual-Domain Self-Denoising Transformer (DuDoST), whose overall framework is illustrated in Fig. 1. DuDoST adopts a ResNet-101 backbone pre-trained on the ImageNet dataset. The multi-scale features produced by the backbone are first fed into a Pixel Decoder for feature fusion and scale alignment. Subsequently, to mitigate the interference of background noise in multi-scale features on the Transformer decoding process while enhancing the model's attention to semantically relevant regions, we introduce Hierarchical Semantic Prior Gating (HSPG). The semantically gated features are then input into the Transformer decoder for query modeling and initial mask prediction. In addition, the Mask-domain Diffusion Denoiser (MaDD), embedded after the Transformer decoder, introduces conditional diffusion modeling into the mask domain by applying random perturbations to the predicted masks and autonomously restoring their structures, thereby endowing the model with stronger structural awareness and self-repair capability during training.

B. Hierarchical Semantic Prior Gating(HSPG)

The overall structure of Hierarchical Semantic Prior Gating (HSPG) is illustrated in Fig. 2. HSPG explicitly introduces semantic priors in the feature space and applies semantic guidance to multi-scale features through hierarchical semantic modeling and consistency gating mechanisms, thereby providing more stable and semantically consistent feature representations for Transformer decoding and subsequent mask prediction.

Specifically, since high-level features typically contain richer semantic information and are less affected by texture noise, this study first selects high-level features from the multi-scale features output by the Pixel Decoder as the input source for semantic priors. Let the multi-scale features produced by the pixel decoder be denoted as:

$$\mathcal{F} = \{P_5, P_4, P_3, P_2\}, \quad P_k \in \mathbb{R}^{C \times H_k \times W_k} \tag{1}$$

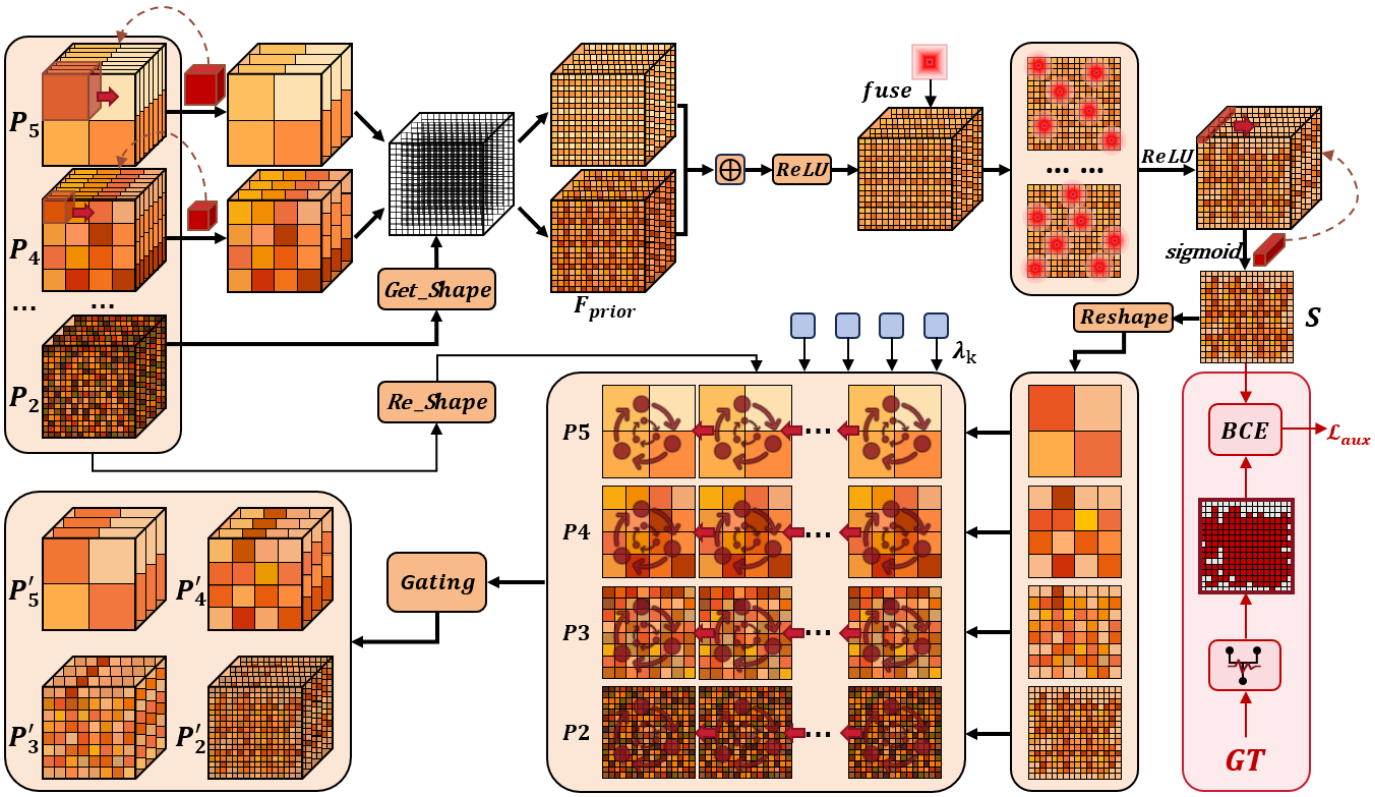

Fig. 2. Structure of the Hierarchical Semantic Prior Gating (HSPG).

Here, C denotes the number of feature channels, while H_k and W_k represent the spatial resolution of features at the k-th scale. P_k denotes the feature map at the k-th scale, where P_2 has the highest spatial resolution and P_5 contains the richest semantic information. In this study, P_5 and P_4 are regarded as high-level semantic features, and channel compression and semantic projection are applied to these two feature levels to unify the representation space across different scales. Subsequently, the projected features are upsampled to the highest resolution level and fused in a pixel-wise manner to generate a hierarchical semantic prior:

$$\mathcal{F}_{prior} = \sum_{k\in\{4,5\}} \mu(Conv_{1\times1}(P_{kv})) \tag{2}$$

Here, $\mu(\cdot)$ denotes the upsampling operation that resizes features to the highest spatial resolution. Subsequently, to suppress discrete noise while enhancing spatial continuity, local modeling is performed on the fused features to generate a spatial semantic prior map:

$$S = \sigma(Conv_{out}(Conv_{3\times3}(\mathcal{F}_{prior}))) \tag{3}$$

Here, $\sigma(\cdot)$ denotes the standard operator for confidence modeling, whose output is normalized to the range $[0, 1]$. $S \in [0, 1]^{1\times H\times W}$ represents the semantic confidence map, where $S_{(x,y)}$ indicates the confidence that the pixel location (x, y) belongs to a semantically relevant region.

After obtaining the semantic prior map S, this prior is consistently injected into feature maps at all scales to enhance semantically relevant regions in the feature space. Specifically, for feature maps at an arbitrary scale $P_k \in \mathcal{F}$, the semantic prior S is first upsampled to the corresponding spatial resolution. Subsequently, an enhanced gating strategy is applied to modulate the features, yielding a new set of multi-scale features:

$$P_k^{'} = P_k \odot (1 + \lambda\mu_k(S)), k \in \{2, 3, 4, 5\} \tag{4}$$

Here, $\mu_k(\cdot)$ denotes the operation that upsamples the semantic prior map to the same spatial resolution as the feature at the k-th level, $\odot$ denotes element-wise multiplication, and λ is the gating strength coefficient. It is worth noting that, in order to enhance the model's responsiveness to semantic foreground regions while ensuring training stability, the proposed design only enhances semantically relevant regions without forcibly suppressing background regions. Finally, the resulting multi-scale features $\{P_5', P_4', P_3', P_2'\}$ are fed into the Transformer decoder for subsequent query modeling and mask prediction.

In addition, the HSPG module is optimized implicitly under an end-to-end framework, where its parameters are updated through backpropagation driven by the loss of the main segmentation task. However, since the semantic prior map S directly participates in feature gating and affects the Transformer decoding stage, its learning process is constrained by the main segmentation objective. Therefore, to further stabilize its learning behavior during the early stages of training,

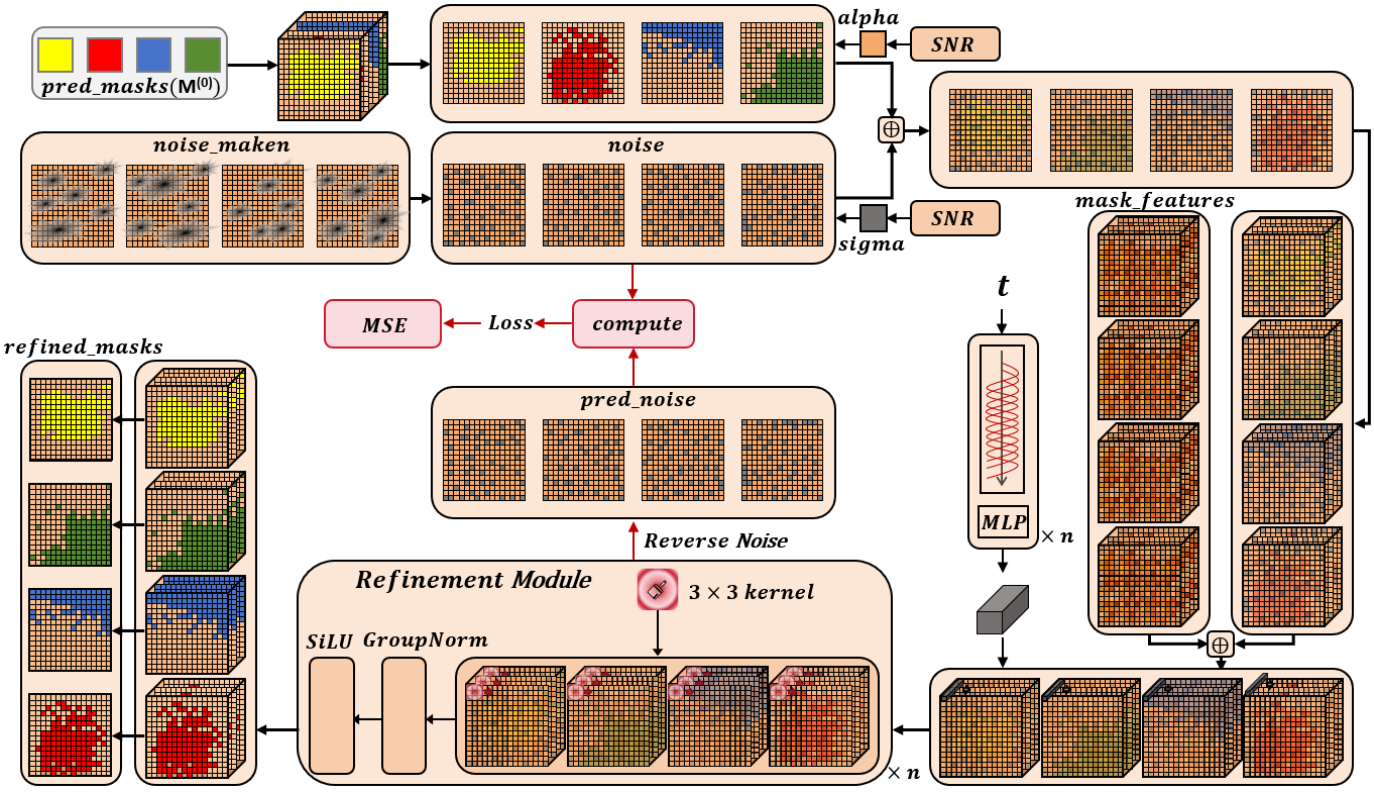

Fig. 3. Structure of the Mask-domain Diffusion Denoiser (MaDD).

we introduce a lightweight auxiliary supervision mechanism. Specifically, a foreground indicator map G is first constructed based on the pixel-level segmentation annotations:

$$G(x,y) = \begin{cases} 1, & if\ GT\_seg(x,y) \neq c_{bg} \\ 0, & otherwise \end{cases} \tag{5}$$

Here, C_{bg} denotes the background class ID. Then, a binary cross-entropy loss is employed to provide auxiliary supervision for the semantic prior, and the final training objective is defined as:

$$\mathcal{L} = \mathcal{L}_{seg} + \alpha BCE(S,G), \alpha \ll 1 \tag{6}$$

Here, $\mathcal{L}_{\text{seg}}$ denotes the main segmentation loss, α is the weighting coefficient of the auxiliary supervision, and $BCE(\cdot)$ denotes the auxiliary binary cross-entropy loss. It should be emphasized that, to avoid introducing excessive interference to the main segmentation task, the auxiliary supervision is used only as a regularization constraint during training and does not introduce additional inference branches or computational overhead, and its weight is set to a relatively small value.Moreover, this module can be implicitly learned through the main task without auxiliary supervision, or further stabilize the training process via lightweight auxiliary supervision when required.

C. Mask-domain Diffusion Denoiser(MaDD)

The overall structure of the Mask-domain Diffusion Denoiser (MaDD) is illustrated in Fig. 3. It explicitly introduces diffusion modeling into the mask domain. Unlike diffusion modeling conducted in the pixel space or feature space, MaDD directly applies random perturbations to the predicted masks and learns to restore their mask structures under the conditional constraint of pixel-level semantic features (*mask_features*), thereby enabling the model to exhibit stronger structural awareness and self-repair capability during training.

Specifically, let the raw mask predictions output by the Transformer decoder (*pred_masks*) be denoted as:

$$M^{(0)} \in R^{B\times Q\times H\times W} \tag{7}$$

Here, B denotes the batch size, Q denotes the number of queries, and $H \times W$ represents the spatial resolution of the masks. Subsequently, for diffusion modeling, the masks are reshaped into $B \times Q$ feature maps, where the mask corresponding to each query is treated as an independent diffusion sample and used as the original input of the diffusion process:

$$x_0 = reshape(M^{(0)}) \in R^{(B\times Q)\times 1\times H\times W} \tag{8}$$

In the mask domain, we adopt a continuous-time diffusion process to apply random perturbations to x_0 in order to simulate noise effects in real scenarios. Specifically, diffusion time steps are first randomly sampled from a uniform distribution, i.e., $t_n \sim \mathcal{U}(0,1)$. Since there are a total of $B \times Q$ feature maps, we index them as $n = 1, 2, \ldots, B \times Q$. Subsequently, according to a predefined signal-to-noise ratio schedule, a

TABLE I
COMPARISON OF DUDOST WITH STATE-OF-THE-ART METHODS ON THE CITYSCAPES DATASET

Method	Roa	Sid	Bui	Wal	Fen	Pol	TLi	TSi	Veg	Ter	Sky	Per	Rid	Car	Tru	Bus	Tra	Mot	Bic	mIoU (%)
BiSeNetV1	98.9	84.9	92.2	57.2	54.8	64.3	70.6	74.0	93.0	71.8	94.8	83.7	64.4	95.1	58.6	72.7	58.2	59.9	71.8	74.8
SANet-75	98.4	84.9	92.2	52.9	56.3	63.2	69.5	73.7	92.9	70.5	94.8	83.4	65.9	95.4	71.3	82.6	74.2	61.2	72.0	76.6
STDC2-Seg75	98.3	84.6	92.1	52.7	56.2	63.7	69.7	73.3	91.7	70.7	95.9	84.8	68.5	95.5	71.0	83.4	75.6	62.7	72.7	77.0
DDRNet-23-slim	98.5	85.6	92.5	52.5	57.2	63.0	71.4	75.2	93.2	71.7	95.4	84.4	68.4	95.5	71.5	83.1	76.6	63.0	73.1	77.4
FeedFormer	98.4	84.4	92.9	54.4	56.6	67.7	75.2	78.1	93.5	71.9	95.6	85.7	69.2	95.5	62.2	81.1	76.3	64.7	74.9	77.9
PSPNet	98.5	85.4	92.8	55.6	59.1	63.3	70.9	75.6	93.4	71.1	95.2	86.4	71.3	95.9	72.3	82.2	72.3	70.4	76.5	78.4
MaskFormer	98.5	85.7	93.0	58.6	58.3	67.3	74.7	77.7	93.5	72.2	95.6	85.3	69.2	95.6	68.1	83.0	78.3	65.9	75.1	78.5
Senformer	98.8	86.4	93.0	53.5	59.8	65.2	75.8	78.9	94.2	72.8	95.6	87.7	71.5	96.1	67.2	80.2	73.4	70.4	77.1	78.8
DeepLab V3+	98.5	85.9	93.1	51.8	59.6	67.8	74.9	78.3	93.5	72.2	95.5	86.8	72.1	95.6	68.9	81.3	74.8	70.2	76.6	79.0
DDRNet-23	98.7	86.6	92.8	47.5	60.5	66.9	75.6	78.5	93.6	72.3	95.5	86.0	71.0	96.0	70.9	86.5	85.9	67.6	76.4	79.4
FRFN	98.9	86.7	93.8	57.4	60.8	66.7	77.2	79.8	93.2	72.3	95.8	86.9	71.8	96.4	73.7	82.5	78.2	71.8	77.6	80.1
Mask2Former	98.8	86.8	93.7	57.4	60.9	67.0	77.1	79.9	93.0	72.4	95.5	86.6	72.1	96.3	72.7	82.6	78.7	71.8	77.5	80.1
PIDNet-L	98.9	87.3	93.9	57.5	61.4	67.3	77.3	80.2	93.4	72.2	95.7	86.3	72.5	96.5	72.9	80.6	78.8	71.8	77.7	80.9
OCRNet	99.0	87.2	94.0	57.6	61.5	67.5	77.0	80.4	93.5	72.4	95.6	86.4	72.7	96.3	73.1	80.9	79.0	70.7	77.9	81.1
FeedFormer-B2	99.1	87.4	94.1	56.9	61.7	67.9	77.2	80.5	93.7	71.9	95.8	86.4	72.9	96.5	72.8	81.1	79.1	70.9	78.0	81.5
DuDoST (Ours)	**99.5**	**88.0**	**94.3**	**57.5**	**61.2**	**67.0**	**78.9**	**81.3**	**94.1**	**72.7**	**96.4**	**86.9**	**73.7**	**96.6**	**72.9**	**81.7**	**79.6**	**71.7**	**78.9**	**82.9**

TABLE II
COMPARISON OF DUDOST WITH STATE-OF-THE-ART METHODS ON THE ADE20K DATASET

Method	mIoU (%)	FLOPs (G)	Params (M)
BiSeNetV1	35.1	15.0	13.3
PIDNet-L	40.5	34.0	37.4
YOSO	44.7	37.3	42.0
OCRNet	45.6	164.8	70.5
SenFormer	46.0	179.0	144.0
MaskFormer	46.7	73.0	60.0
Mask2Former	47.7	90.0	63.0
FeedFormer-B2	46.0	42.7	29.1
DuDoST (Ours)	**50.5**	103.0	64.7

preset function $\gamma(t)$ is used to compute the corresponding signal and noise coefficients:

$$\gamma(t) = \log \frac{\alpha^2(t)}{\sigma^2(t)} \tag{9}$$

$$\alpha(t) = \sqrt{\sigma(\gamma(t))}, \qquad \sigma(t) = \sqrt{\sigma(-\gamma(t))} \tag{10}$$

Here, $\alpha(t)$ controls the retention ratio of the original mask, while $\sigma(t)$ controls the influence of noise on the original feature map. $\gamma(t)$ is a predefined function that maps the corruption level to the corresponding signal-to-noise ratio. Based on this formulation, a perturbed mask representation is then constructed as:

$$x(t) = \alpha(t)x_0 + \sigma(t)\epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{11}$$

Here, ϵ is a Gaussian noise map with the same shape as x_0, sampled from a standard multivariate Gaussian distribution with zero mean and identity covariance $\mathcal{N}(0, I)$, where I denotes the identity covariance matrix, ensuring pixel-wise independent and identically distributed noise. It should be emphasized that diffusion noise is injected only into the predicted masks, while the pixel-level semantic features, serving as a stable semantic description of the input image, remain unaffected and are used as conditional constraints for the subsequent denoising process.

Next, to recover the original structure of the masks after structural perturbation, we further design a conditional denoising network. This network takes the previously perturbed mask $x(t)$ as input and outputs the network-predicted clean mask $\hat{x}_0$:

$$\hat{x}_0 = \mathcal{D}_\theta(x(t), F, \phi(t)) \tag{12}$$

Here, $D_\theta(\cdot)$ denotes a learnable conditional denoising network. $F \in \mathbb{R}^{(B\cdot Q)\times C\times H\times W}$ represents the pixel-level semantic features (*mask_features*), which are produced by the pixel decoder and aligned to the query dimension, serving as stable semantic references when the masks are perturbed. $\phi(t)$ denotes the diffusion time embedding process, which maps the time step t to a high-dimensional vector via sinusoidal positional encoding followed by a multilayer perceptron (MLP), explicitly informing the network of the severity of noise corruption in the current mask.

In addition, to avoid issues such as blurred predictions caused by directly regressing the masks, we leverage the analytical relationship of the forward diffusion process to infer the noise estimation $\hat{\epsilon}_\theta$, and train the model using the noise regression objective commonly adopted in standard diffusion models:

$$\hat{\epsilon}_\theta(t) = \frac{x(t) - \alpha(t)\hat{x}_0}{\sigma(t)} \tag{13}$$

$$\mathcal{L}_{MaDD} = \mathbb{E}_{t\sim\mathcal{U}(0,1),\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[\|\hat{\epsilon}_\theta(t) - \epsilon\|_2^2\right] \tag{14}$$

In this way, when subjected to different levels of noise corruption, the network can accurately model the relationship between mask structures and noise, thereby learning more robust structural recovery capability.

Finally, during training, the denoised masks $\hat{x}_0$ produced by MaDD are reshaped as $M \in \mathbb{R}^{B\times Q\times H\times W}$, and directly replace the original mask predictions for subsequent Hungarian

TABLE III
ABLATION STUDIES OF HSPG ON THE CITYSCAPES DATASET

Method	Source	Gating	mIoU (%)	FLOPs (G)	Params (M)
Baseline	–	–	80.1	628.0	67.0
+HSPG	P_5	–	80.9	666.8	67.2
+HSPG	P_5, P_4	–	81.3	667.3	67.2
+HSPG	P_4, P_3	–	80.8	669.4	67.2
+HSPG	P_3, P_2	–	80.4	677.4	67.2
+HSPG	P_5, P_4	Hard $(S \odot P)$	81.0	667.4	67.2
+HSPG	P_5, P_4	Additive $(P + \lambda S)$	81.2	667.4	67.2
+HSPG	P_5, P_4	Enhancement $\big(P \odot (1 + \lambda S)\big)$	81.5	667.4	67.2
+HSPG + Aux Loss	P_5, P_4	Enhancement $\big(P \odot (1 + \lambda S)\big)$	81.6	667.4	67.2

TABLE IV
ABLATION STUDY ON THE NUMBER OF TIMESTEPS IN MADD ON THE ADE20K DATASET

Method	Timesteps	mIoU (%)	FLOPs (G)	Params (M)
Baseline	0	47.7	90.0	63.0
+ MaDD	1	49.0	99.0	64.5
+ MaDD	2	49.4	118.0	64.5
+ MaDD	3	49.6	126.0	64.5

TABLE V
JOINT ABLATION STUDY OF HSPG AND MADD ON CITYSCAPES AND ADE20K DATASETS

Method	mIoU (%)	FLOPs (G)	Params (M)
Cityscapes			
Baseline	80.1	628.0	67.0
+HSPG	81.6	667.4	67.2
+MaDD	81.7	684.5	68.5
+HSPG + MaDD	82.9	724.0	68.7
ADE20K			
Baseline	47.7	90.0	63.0
+HSPG	48.9	94.9	63.2
+MaDD	49.0	99.0	64.5
+HSPG + MaDD	50.5	103.0	64.7

matching and segmentation loss computation. The final joint optimization loss is then formulated as:

$$\mathcal{L} = \mathcal{L}_{seg}(M) + \lambda \mathcal{L}_{MaDD} \tag{15}$$

Here, $\mathcal{L}_{\text{seg}}$ denotes the original classification and mask loss in Mask2Former, and λ is the weighting coefficient. Through this design, MaDD is no longer an independent auxiliary branch, but is instead deeply coupled with the mask prediction pipeline, enabling the model to automatically bias toward generating segmentation masks with more stable structures, more consistent boundaries, and easier structural recovery during end-to-end training.

IV. EXPERIMENTS

A. Datasets

The Cityscapes dataset is a semantic image dataset focusing on urban street scenes, covering street scenarios from 50 different cities under favorable weather conditions. It contains a total of 5,000 finely annotated images and more than 30 semantic categories. Among them, 19 object categories are provided with dense pixel-level annotations. The dataset is split into 2,975 training images, 500 validation images, and 1,525 test images. The image resolution is 1024×2048.

ADE20K is a scene parsing dataset that contains 25,000 images of complex daily scenes, covering 150 semantic categories and a wide variety of objects in natural environments. It includes 20,210 training images, 2,000 validation images, and 3,352 test images. On average, each image contains 19.5 instances and 10.5 object categories.

B. Implementation Details

All training and evaluation of the proposed DuDoSeg are implemented using the PyTorch framework on a single NVIDIA A100 GPU with 40GB memory. The model adopts Mask2Former [7] as the baseline, and the overall experimental procedure follows its standard training paradigm. The Transformer decoder consists of 10 decoding layers, each equipped with 8 attention heads, and the hidden feature dimension is set to 256. The AdamW optimizer is used with a base learning rate of 1×10^{-4} and a weight decay coefficient of 0.05. The learning rate of the backbone network is scaled by a factor of 0.1, and L2 gradient norm clipping with a maximum value of 1.0 is applied. The auxiliary supervision weight coefficient α for HSPG ($\alpha \ll 1$) is used only as a regularization constraint and is linearly warmed up during the early stage of training. The MaDD module introduces random diffusion noise to the predicted masks only during training and learns the corresponding recovery process, while during inference it performs only a single forward denoising prediction.

C. Comparison With State-of-the-art Methods

a) Comparisons on Cityscapes: As shown in Table I, the proposed DuDoST is quantitatively compared with various mainstream semantic segmentation methods on the Cityscapes dataset in terms of 19 semantic categories and the overall mIoU metric. From an overall performance perspective, DuDoST achieves an mIoU of 82.9%, demonstrating a clear performance advantage over other advanced methods. From a category-level analysis, DuDoST not only exhibits strong modeling capability for large-object structures and semantics, but also attains outstanding accuracy on small objects and fine-grained categories. In summary, the results in Table I

fully validate the effectiveness and stability of DuDoST on the Cityscapes dataset. It not only achieves state-of-the-art performance in terms of overall mIoU, but also reaches optimal or near-optimal results on most semantic categories, indicating that it achieves a better balance between global modeling and overall performance, while further demonstrating robustness in modeling local details and small-scale objects.

b) Comparisons on ADE20K: As shown in Table II, DuDoST is compared with several representative semantic segmentation methods on the ADE20K dataset in terms of both performance and computational complexity. In terms of segmentation accuracy, DuDoST achieves the highest mIoU of 50.5% on ADE20K, significantly outperforming all compared methods. Regarding computational complexity, DuDoST maintains a favorable balance between efficiency and performance while preserving high accuracy. Overall, on the highly complex ADE20K semantic segmentation dataset, DuDoST not only attains state-of-the-art accuracy but also maintains reasonable control over model size and computational cost. This further validates the generalization capability and practical value of DuDoST in multi-class and fine-grained semantic understanding scenarios.

D. Ablation Studies and Analysis

a) Ablation Studies of HSPG: As shown in Table III, systematic ablation experiments of the proposed HSPG module are conducted on the Cityscapes dataset from multiple aspects. First, after integrating the HSPG module into the backbone network, the model performance is improved. Specifically, using the combination of P_5 and P_4 as senior semantic priors achieves the highest mIoU of 81.3%, outperforming configurations that rely on a single high-level or lower-level feature. On this basis, enhanced gating enables the model to achieve the best mIoU of 81.5% while maintaining stable computational complexity, showing advantages over direct hard gating and additive fusion. Furthermore, after introducing auxiliary supervision (Aux Loss), the model performance is further improved to 81.6% mIoU. Meanwhile, the FLOPs and parameter count remain unchanged (667.4 GFLOPs / 67.2M Params), which fully demonstrates that the auxiliary supervision does not introduce additional inference branches. Overall, the ablation results comprehensively validate the effectiveness of each component of the HSPG module, indicating that it can significantly improve segmentation performance with only negligible increases in parameters and computational cost.

b) Ablation Studies of MaDD: As shown in Table IV, ablation experiments are conducted on the ADE20K dataset to evaluate MaDD under different inference timestep settings. In terms of segmentation accuracy, when MaDD is introduced with single-step sampling (*timesteps*=1), the model achieves an mIoU of 49.0%. However, as the number of inference timesteps increases, the performance gain noticeably diminishes after *timesteps*=2, exhibiting a phenomenon of diminishing marginal returns. In terms of computational complexity, the number of parameters remains unchanged (approximately 64.5M) after introducing MaDD, indicating that MaDD does not introduce additional parameters with increasing sampling steps. Nevertheless, the computational cost during inference grows multiplicatively with the number of timesteps. Therefore, considering the trade-off between segmentation accuracy and computational overhead, this paper adopts *timesteps*=1 as the default inference configuration.

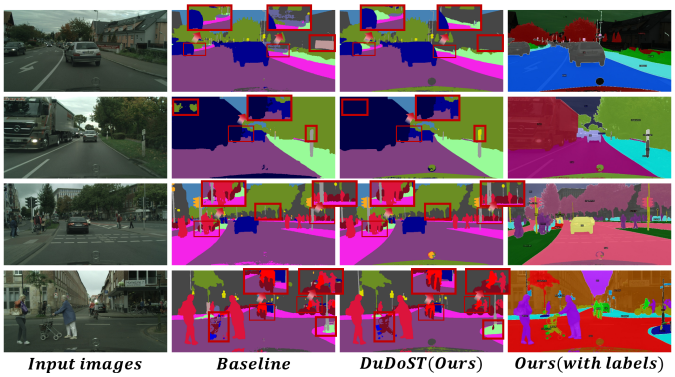

Fig. 4. Visual comparison results of DuDoST and Baseline methods on the Cityscapes dataset.

c) Joint Ablation Study of MaDD and HSPG: As shown in Table V, joint ablation experiments of the two key modules, HSPG and MaDD, are conducted on the Cityscapes and ADE20K datasets. When only HSPG is introduced, the performance on the Cityscapes and ADE20K datasets is improved to 81.6% and 48.9% mIoU, respectively, verifying its effectiveness in enhancing the overall model performance. When only MaDD is introduced, the model performance also shows a significant improvement on both datasets, indicating its substantial contribution to segmentation accuracy. When HSPG and MaDD are jointly applied, the model achieves the best mIoU of 82.9% and 50.5% on the two datasets, respectively, demonstrating that the two modules exhibit complementary effects rather than a simple additive relationship, and that enabling both leads to more comprehensive structural modeling capability. Meanwhile, the model parameters and computational cost remain within a reasonable range. Overall, the ablation results in Table V fully validate the individual effectiveness of HSPG and MaDD, as well as their clear complementary advantages in tasks involving complex semantic distributions and fine-grained structural modeling.

E. Visualization Analysis

As illustrated in Fig. 4, to further validate the effectiveness of DuDoST, a qualitative comparison with the baseline method Mask2Former is conducted on the Cityscapes dataset. From the overall visual results, DuDoST is able to generate smoother and more semantically consistent segmentation outputs. It demonstrates stronger robustness on small-scale objects such as pedestrians and bicycles, as well as in complex occlusion scenarios, effectively reducing mis-segmentation and semantic fragmentation. Meanwhile, DuDoST preserves clearer and more complete structural boundaries in large-scale regions such as roads and sky, avoiding unnecessary noisy predictions.

These qualitative results further confirm the effectiveness of the proposed method in suppressing prediction noise, refining object boundaries, and enhancing multi-category semantic understanding.

V. Conclusion

This paper proposes a dual-domain collaborative segmentation framework, termed DuDoST, to address noise interference and semantic instability in Transformer-based semantic segmentation. The proposed framework systematically enhances the stability and consistency of segmentation predictions from two complementary perspectives, namely the feature space and the mask space. In the feature modeling stage, HSPG is introduced to effectively suppress background noise and stabilize the input distribution of the Transformer decoder. In the mask prediction stage, MaDD is further incorporated to improve boundary quality and overall shape rationality.Extensive experiments conducted on the Cityscapes and ADE20K datasets demonstrate that the proposed model achieves more robust and semantically consistent segmentation results without significantly increasing inference complexity, and exhibits improved generalization ability and practical value under complex scenes, small-object conditions, and fine-grained structural scenarios.

Looking forward, future work may extend the proposed dual-domain collaborative framework to instance segmentation or panoptic segmentation, exploring a unified dual-domain adaptive image segmentation framework, or investigate more lightweight noise injection and denoising strategies to further reduce model complexity. We believe that DuDoST provides important insights for Transformer-based semantic segmentation models.

Acknowledgments

The authors thank the Natural Science Foundation of Hebei Province, China (Project No. F2024201012) for financial support, and the High-Performance Computing Center of Hebei University for providing computational resources.

References

[1] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, and Y. Fu, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 6881–6890.

[2] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 12179–12188.

[3] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 7262–7272.

[4] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: simple and efficient design for semantic segmentation with transformers," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021.

[5] B. Cheng, A. G. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021.

[6] Y. Zhang, H. Zhang, X. Liu, and J. Yang, "K-Net: towards unified image segmentation," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021.

[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.

[8] L. Hoyer, D. Dai, H. Wang, and L. Van Gool, "DAFormer: improving network architectures and training strategies for domain-adaptive semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 9924–9935.

[9] J. Ding, Y. Li, J. Wang, and R. Ji, "MP-Former: mask prediction transformer for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 421–437.

[10] J. Jain, J. Li, M. Najibi, A. G. Schwing, and A. Kirillov, "OneFormer: one transformer to rule universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 2989–2998.

[11] Y. Zhang, D. Gong, Y. Liu, and M. Zhang, "Progressive domain adaptation for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 12577–12586.

[12] Z. Fang, H. Zhang, J. Yang, and Y. Liu, "UniGS: unified generative segmentation via denoising diffusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

[13] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2881–2890.

[14] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

[15] H. Zhang, J. Xiong, L. Wang, J. Chen, and S. C. Zhu, "Context encoding for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7151–7160.

[16] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 1971–1980.

[17] Y. Yuan and J. Wang, "OCNet: object context network for scene parsing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 2378–2387.

[18] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 173–190.

[19] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 603–612.

[20] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, "Asymmetric non-local neural networks for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 593–602.

[21] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 3146–3154.

[22] Y. Zhang, H. Zhang, X. Liu, and J. Yang, "Part-aware context aggregation for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 635–651.

[23] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2980–2988.

[24] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with gaussian edge potentials," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2011, pp. 109–117.

[25] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 5168–5177.

[26] A. Kirillov, Y. Wu, K. He, and R. Girshick, "PointRend: image segmentation as rendering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 9799–9808.

[27] G. Ke, Y. Chen, Z. Huang, S. Qiu, and Y. Xu, "SegFix: model-agnostic boundary refinement for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 489–506.

[28] H. Ding, X. Jiang, B. Shuai, A. Qun Liu, and G. Wang, "Boundary-preserving refinement for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 308–324.