Journal of Multimedia Information System VOL. 13, NO. 2, June 2026 (pp. 43-52): ISSN 2383-7632 (Online)
https://doi.org/10.33851/JMIS.2026.13.2.43

Single Image Depth Estimation Based on CLIP Language Prior Injection

Sheng Yu[1], Jun Miao[1*], Jianqiao Wang[1]

Abstract

Single-image depth estimation models perform well on data within their training distributions but exhibit insufficient generalization and robustness when confronted with out-of-distribution data or complex scenes. Existing approaches are largely confined within the visual modality, lacking high-level semantic understanding and struggling with challenges like occlusion and texture loss. To address this, we introduce the Contrastive Language and Image Pre-training (CLIP) model, which achieves strong open-world generalization capabilities through contrastive learning on 200 million text-image pairs. However, a gap exists between CLIP's scene-level representations and the pixel-level geometric regression required for monocular depth estimation. To bridge this gap, we design a Text-guided Reverse Cross-Attention (TRCA) module to simulate the relationship between text and pixels, enabling fine-grained semantic enhancement from text to pixels. Concurrently, we propose a Text-Guided Depth Bins Center Predict (TBCP) module that utilizes gated attention to filter scene structural semantic features, transforming continuous depth regression into semantically guided discrete distribution estimation. Experimental results on the NYU Depth v2 dataset show that our method achieves a 3.5\% reduction in absolute relative error (Abs Rel) and a 5\% reduction in root mean square error (RMSE) compared to PixelFormer. Furthermore, evaluations across four distinct test datasets confirm that our model exhibits superior generalization capability and enhanced robustness for depth estimation in complex scenes.

Key Words: Single Image, Depth Estimation, CLIP, Dense Prediction, Attention Mechanism.

I. INTRODUCTION

Single Image Depth Estimation (SIDE) has garnered significant attention due to its low cost, lack of additional sensors, and ease of deployment. Unlike traditional depth sensing technologies such as LiDAR or stereo vision, SIDE requires only a single RGB image as input, offering exceptionally broad application prospects spanning applications such as autonomous driving [1], virtual reality [2], and robot navigation [3]. This versatility presents significant challenges: achieving exceptional generalization capabilities to effectively handle the diversity and complexity of diverse application scenarios. However, variations in scene layout, depth distribution, and lighting conditions render this an exceptionally challenging task.

In recent years, to enhance the generalization capabilities of monocular depth estimation models in unseen domains, primary research approaches have been categorized into two types: data-driven methods [4-6] and model-driven methods [7-9]. The former relies on large-scale, meticulously annotated image-depth pairs to map images to depths, requiring time-consuming and labor-intensive data collection and training processes. In contrast, model-driven approaches aim to leverage rich prior knowledge embedded in pre-trained models to enhance generalization performance. For instance, PixelFormer [9] incorporates skip attention mechanisms within the Transformer [10] architecture based on attention, effectively integrating high-resolution local texture features from the encoder with global contextual information from the decoder. This facilitates long-range propagation of semantic information, significantly mitigating label confusion in pixel-level depth estimation and enhancing model robustness. However, this approach still exhibits noticeable bias in depth predictions under complex conditions such as distant views and strong illumination. This limitation primarily stems from such models' over-reliance on intra-modal visual features and insufficient understanding of high-level semantic information, resulting in inadequate adaptability to challenging scenarios involving occlusions or texture deficiencies.

To mitigate this constraint, this paper introduces the contrastive language-image pre-training model CLIP [11] as external semantic prior. CLIP, pre-trained on 200 million

Manuscript received January 23, 2026; Revised February 25, 2026; Accepted March 18, 2026. (ID No. JMIS-26M-01-003)
Corresponding Author (*): Jun Miao, +86-180-7912-3391, miaojun@nchu.edu.cn
[1]Department of Research & Education, Jiangxi Key Laboratory of Image Processing and Pattern Recognition, Nanchang Hangkong University, Nanchang, China, yu_depth2026@outlook.com, miaojun@nchu.edu.cn, 1586133556@qq.com

image-text pairs, possesses robust cross-modal alignment and cross-domain generalization capabilities. Its rich high-level semantic information effectively supplements missing visual representations in complex environments, enhancing the model's semantic perception and depth inference capabilities in challenging scenes. However, a significant granularity gap exists between its scene-level representations and the pixel-level geometric outputs required for depth estimation. To address this, this paper proposes a framework bridging CLIP's macro-level semantics with pixel-level geometric understanding. At the feature representation level, we introduce a Text-guided Reverse Cross-Attention matching method. By modeling the relationship between textual descriptions and pixel-level features, this approach adaptively injects scene-level semantic priors into the pixel-level depth estimation task, thereby enhancing the semantic consistency of monocular depth estimation in complex scenes. At the scene understanding level, we introduce a Text-Guided Depth Bins Center Predict module. This module discretizes continuous depth into multiple intervals via a depth binning adapter and employs a gated attention mechanism to filter scene structural information from textual features. This achieves an adaptive mapping from semantic priors to depth distributions, thereby enhancing model robustness in complex scenes. The main contributions of this paper are summarized as follows:

(1) We introduce a novel paradigm that adapts scene-level textual descriptions to pixel-level depth estimation, effectively leveraging semantic priors from pre-trained vision-language models;
(2) By introducing two core modules-TRCA and TBCP working synergistically, we significantly enhance the generalization capability of the SIDE model in complex scenes;
(3) Experiments on NYU Depth v2 show our method reduces Abs Rel and RMSE by 3.5% and 5%, respectively, versus PixelFormer, with further cross-dataset tests confirming superior generalization.

II. RELATED WORK

2.1. Single Image Depth Estimation

Single Image Depth Estimation (SIDE) has achieved notable progress on standard benchmarks [12]. Following the pioneering CNN-based work of Eigen et al. [13], subsequent research has evolved along three main avenues: (A) Architecture advancements, including the use of residual networks [14], multi-scale fusion strategies [15], Vision Transformers [16], and diffusion models [17]. (B) Improved optimization, such as classification regression paradigms [18] and geometric constraints [19]. and (C) Multi-task learning with auxiliary cues like surface normal [20] or semantics. While effective on indistribution data, these methods often lack the robustness required for real-world applications under diverse conditions. This has shifted the research focus towards generalizable SIDE.

Current efforts for generalization are primarily two-fold. The first, data-driven, relies on large-scale, often web-collected, image-depth pairs. Methods like MiDaS [4] predict affine-invariant relative depth, enabling training on mixed datasets. Omnidata [5] and Depth Anything [12] expanded this concept with massive datasets for zero-shot transfer. Others, like ZoeDepth [6], fine-tune such relative depth models for metric output. While powerful, this paradigm demands significant data and compute. The second, model-driven, leverages strong priors from models pre-trained on even larger datasets. Notably, vision-language models like CLIP-trained on 200 million image-text pairs offer rich, semantically-aligned cross-modal representations. Harnessing these priors for SIDE presents a promising path for improved generalization without exhaustive data collection.

2.2. Large-Scale Visual-Language Models

The success of CLIP has spurred its application to various downstream vision tasks, including monocular depth estimation. For instance, Kim et al. [22] proposed a tuning-free method that realigns CLIP's internal representations via engineered text prompts to elicit depth-aware priors. Chatterjee et al. [23] systematically studied the robustness of linguistic guidance in low-level vision, establishing an evaluation benchmark and a multi-level language dataset for depth estimation. However, these approaches primarily exploit CLIP's global image-text matching capability, falling short of achieving pixel-level semantic understanding. This limits their potential in dense prediction tasks like depth estimation, where fine-grained spatial alignment is crucial. Bridging this gap-from global matching to pixel-wise adaptation-remains an open challenge.

To this end, we propose a novel text-to-pixel framework that leverages CLIP's knowledge in a fine-grained manner. Our model consists of two core components: (1) the TRCA module, which establishes pixel-wise semantic correspondence between image features and text descriptions, enabling fine-grained semantic enhancement from text to pixels. and (2) the TBCP module, which discretizes continuous depth into learnable bins and uses gated attention to filter text-based structural cues, thereby mapping semantic priors to depth distributions adaptively. Together, these modules significantly improve the robustness of depth estimation in complex scenes.

Journal of Multimedia Information System VOL. 13, NO. 2, June 2026 (pp. 43-52): ISSN 2383-7632 (Online)
https://doi.org/10.33851/JMIS.2026.13.2.43

III. METHODOLOGY

3.1. Network Architecture

Conventional SIDE methods rely solely on visual features, which often leads to significant errors in textureless or occluded regions. To address this, we enhance the Pixelformer architecture with cross-modal priors from CLIP, formulating depth estimation as a classification-regression task. As illustrated in Fig. 1, our framework operates as follows.

The input image I is processed by a Swin Transformer backbone [24] to extract multi-scale visual features {E_1, E_2, E_3, E_4} at 1/4, 1/8, 1/16, and 1/32 resolutions, capturing global context via window-based self-attention.

In parallel, a human-written scene description T is tokenized and encoded by the CLIP Text Encoder (CLIP-ViT-L/14) to produce a 1024-dimensional global text embedding F_{text}. The core of our method consists of two novel modules that fuse these modalities. First, the TRCA module uses F_{text} to semantically enhance the multi-scale visual features. The enriched features are then passed to a pixel query initializer (PQI), which aggregates global context via multi-scale average pooling to generate initial pixel queries. Second, the TBCP module discretizes the continuous depth range into n_{bins} adaptive bins. It utilizes a gated attention mechanism to filter structural cues from F_{text}, producing a per-pixel probability distribution over the bins. The final depth d_i for pixel is obtained by a weighted sum of the bin center values using this distribution.

3.2. TRCA Module

To effectively fuse linguistic and visual features while preserving visual details, we design the TRCA module, as illustrated in Fig. 2(A) as an interaction module between linguistic and visual features. Since E_4 and F_{text} differ in spatial dimensions and channel counts, a convolution operation reduces E_4 dimensions to match F_{text} channel count. This channel alignment preserves E_4 spatial dimensions, preventing information loss from reduced spatial resolution. Next, the query vector Q from F_{text} is dot-multiplied with the key vector K from E_4 to obtain the self-similarity score matrix S. The similarity score values determine the weights assigned to the value vector V from E_4 in the final output. The weighted sum of S values and vector V is then mapped back to the spatial dimension matching the original visual features via a multilayer perceptron (MLP).

$$F_{fusion} = \mathrm{SV}=\mathrm{Softmax}\left(\frac{\mathrm{QK}^T}{\sqrt{d}}\right)V, \tag{1}$$

where $\sqrt{d}$ denotes the dimension of the query and key vectors, and Q, K, V are all $N \times 2C$ in size, with $N = h \times w$ representing the sequence length.

3.3. TBCP Module

The proposed TBCP module is illustrated in Fig. 2(B). Structurally, it retains the core idea of adaptive prediction of depth interval centers from the Pixelformer method while functionally introducing a gated attention mechanism. This mechanism effectively filters scene structural information from textual features, thereby adaptively mapping semantic priors to the depth distribution. Similar to the TRCA module, Q_4 and text features F_{text} are first mapped to the same spatial dimension via convolution, as

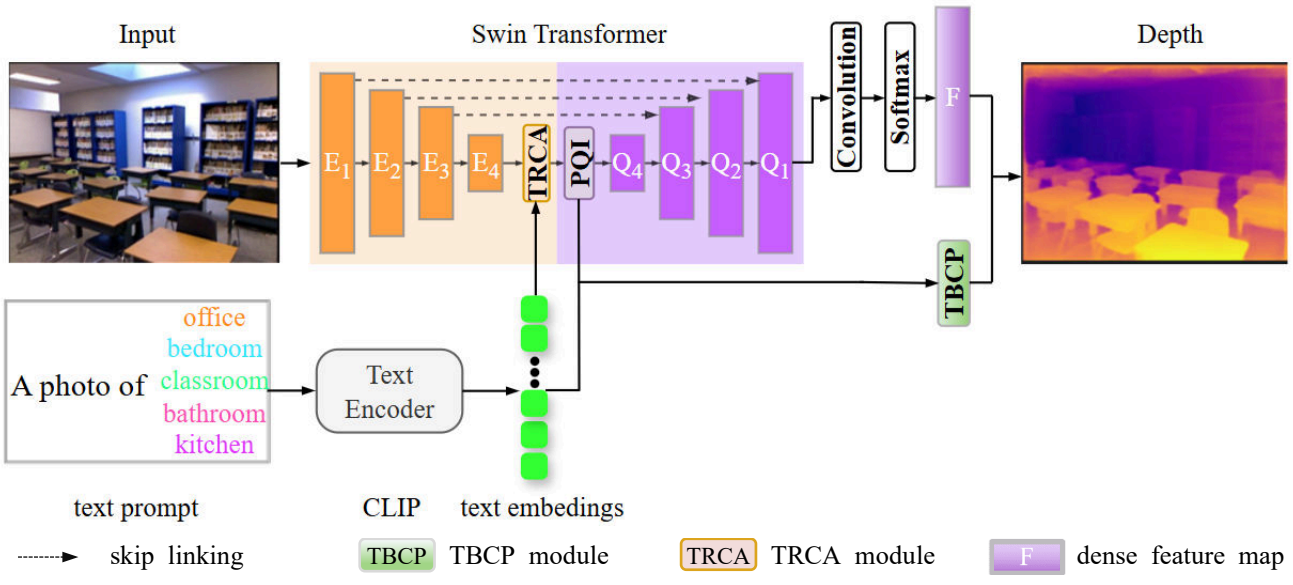
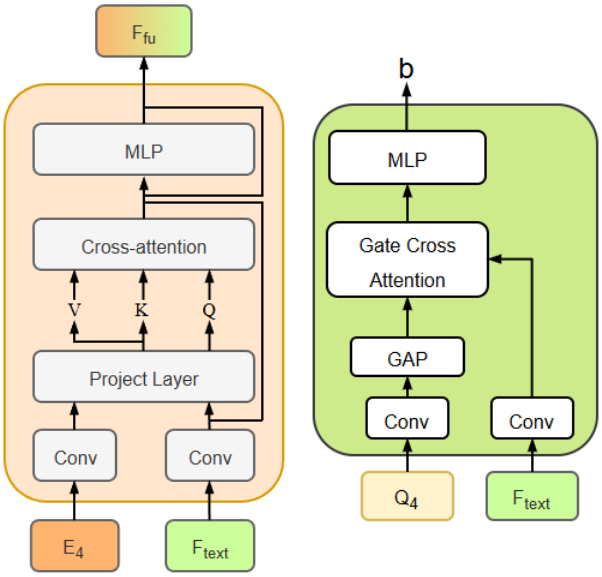


Fig. 1. Flowchart of the proposed network model.



(A) TRCA (B) TBCP

Fig. 2. Module structure diagram.

shown in Fig. 2(A). Global average pooling is then applied to pixel query Q4 to extract a global feature vector integrating visual-semantic information. Subsequently, adaptive gating is employed to fuse this with F_{text}:

$$F_c = W_g \cdot Concat(F_{gap}, F_{text}), \tag{2}$$

$$F_{fu} = \sigma F_c \odot F_{gap} + (1-\sigma) F_c \odot F_{text}, \tag{3}$$

where W_g denotes the gating weight matrix, F_{gap} represents the globally averaged pooling-processed visual feature vector, F_{text} is the global text embedding feature vector, σ is the Sigmoid activation function, and $\odot$ denotes element-wise multiplication.

We generate fused features that simultaneously reflect scene visual characteristics and linguistic attributes. subsequently predict the width distribution b across n_{bins} of depth intervals after processing through a 3-layer MLP.

$$b = MLP(Gap(Q_4, F_{text})), \tag{4}$$

where $Gap(Q_4, F_{text})$ denotes the language-visual fused features processed by global average pooling.

The center of each depth interval is calculated within the dynamically adjusted depth range $[d_{min}(T), d_{max}(T)]$ guided by linguistic context, ensuring the intervals align with both the image's actual depth characteristics and the semantic meaning described in the text. Finally, the center of the *i*-th depth interval is computed as:

$$c(b_i) = d_{min} + (d_{max} - d_{min}) \left(\frac{b_i}{2} + \sum_{j=1}^{i-1} b_j \right), \quad i \in \{1, \ldots, n_{bins}\}, \tag{5}$$

where n_{bins} denotes the number of depth bins, set to 256 in this paper following the configuration in [1]. d_{min} represents the minimum depth value, and d_{max} denotes the maximum depth value.

Finally, at each pixel location, the final depth value $\tilde{b}$ is computed as a linear combination of the Softmax score at that pixel and the depth interval $c(b_i)$, as shown in equation (6):

$$\tilde{b} = \sum_{k=1}^{n_{bins}} c(b_k) p_{k,} \tag{6}$$

where p_k denotes the Softmax score of the pixel within the depth interval.

IV. EXPERIMENTAL RESULTS

4.1. NYU Depth v2 Dataset

NYU v2 [26] is an indoor dataset comprising 120,000 RGB images and corresponding depth pairs of size 480×640. These depth pairs were extracted from video sequences captured using a Microsoft Kinect across 464 indoor scenes. This paper evaluates the algorithm using the official training and test set partition: 249 scenes comprising 50,000 images for training and 654 images for testing. We adopt the centroid segmentation proposed by Eigen31, with a depth map upper bound of 10 meters. The network outputs depth predictions at 120×160 resolution, which we upsampled by a factor of 4 during training and testing to match ground truth resolution.

4.2. Experimental Evaluation Methods

For quantitative assessment, we employ four metrics introduced by Eigen [13], which are most widely used in evaluating monocular depth estimation performance. Among these, Abs Rel denotes the absolute relative error at each pixel location, measuring the discrepancy between predicted and ground truth depths. RMSE represents the root mean square error between predicted and ground truth depths. $\log_{10}$ is the average of the absolute values of the logarithmic differences between the predicted and true depths for each pixel. Acc is the proportion of pixels where the error bet-ween the predicted and true depths is below a certain threshold. They are defined as follows:

$$\text{Abs Rel} = \sum_{i=1}^{N} \frac{|d_i - d_i^*|}{d_i^*}, \tag{7}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} |d_i - d_i^*|^2}, \tag{8}$$

$$log_{10} = \frac{1}{N}\sum \|log_{10} d_i^* - log_{10} d_i\|, \tag{9}$$

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N} 1\left(\max\frac{d_i}{d_i^*}, \frac{d_i^*}{d^i} < \text{thr}\right) \times 100, \tag{10}$$

where N denotes the total number of pixels. d_i represents the predicted depth value for the i-th pixel. d_i^* denotes the ground truth depth label for the i-th pixel. thr is the error threshold, typically set to 1.25, 1.25^2 or 1.25^3. This paper introduces parameters (Parameters, Params) and floating-point operations (FLOPs) to evaluate the portability and computational complexity of the network model.

4.3. Comparative Experiments

To enhance the credibility of the framework results, comparative experiments were conducted. Models were evaluated under identical experimental conditions on the NYU-Depth v2 dataset, including Eigen et al. [13], Yin et al. [19], BTS [27], Naderi [8], TransDepth [29], DPT [21], PackNet-SAN [30], Adabins [18], P3Depth [25], NeWCRFs [5], and PixelFormer [9]. Through comparative experiments, we analyze performance on the NYU-Depth v2 dataset to better evaluate the strengths and weaknesses of the proposed framework. Experimental results are shown in Table 1.

As shown in Table 1 compares our method with recent state-of-the-art approaches, including Eigen, BTS, AdaBins, NeWCRFs, and PixelFormer. Our method achieves the best performance on key metrics: an Abs Rel of 0.088 and an RMSE of 0.319. This represents a 3.5% reduction in Abs Rel and a 5% reduction in RMSE compared to the strong baseline PixelFormer. The improvement in accuracy thresholds (δ_1, δ_2) further confirms the effectiveness of integrating cross-modal semantic prior.

To assess generalization, we directly evaluate models rained on NYU Depth v2 on the unseen SUN RGB-D dataset without fine-tuning. As shown in Table 2, our method outperforms all compared methods, including PixelFormer. This demonstrates that the semantic priors from CLIP significantly enhance the model's robustness to domain shifts. As shown in Fig. 3, compared to PixelFormer, the proposed method demonstrates greater robustness in complex indoor scenes. In the first row of high-glare scenes, depth estimation within the boxed area—covering reflective glass and surrounding environments—is more accurate with clearer hierarchical structure. In the second row of high-brightness metallic scenes, the prediction of highly reflective areas indicated by arrows is more stable. In the third row of large-scale indoor scenes, the edges of distant walls are rendered more clearly, with overall depth hierarchy appearing more distinct.

Fig. 4 further visualizes results on four distinct indoor test sets, showing that our framework consistently produces more geometrically coherent and robust depth estimates in complex, unseen environments, thanks to the semantically guided feature enhancement and depth discretization mechanisms.

Table 1. Comparison experiment results on the NYU-Depth v2 dataset.

Method	Venue	Abs Rel ↓	RMSE ↓	$\log_{10}$ ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Eigen et al. [13]	NIPS'14	0.158	0.641	-	0.769	0.950	0.988
DAV [28]	ECCV'20	0.108	0.412	-	0.882	0.980	0.996
TransDepth [29]	ICCV'21	0.106	0.365	0.045	0.900	0.983	0.996
DPT [21]	ICCV'21	0.110	0.367	0.045	0.904	0.988	**0.998**
PackNet-SAN [30]	CVPR'21	0.106	0.393	-	0.892	0.979	0.995
Adabins et al. [18]	CVPR'21	0.103	0.364	0.044	0.903	0.984	<u>0.997</u>
Naderi et al. [8]	WACV'22	0.097	0.444	0.042	0.897	0.982	0.996
Lee et al. [27]	WACV'22	0.107	0.373	0.046	0.893	0.985	<u>0.997</u>
P3Depth [25]	CVPR'22	0.104	0.356	0.043	0.898	0.981	0.996
NeWCRFs [5]	CVPR'22	0.095	0.334	<u>0.041</u>	0.922	**0.992**	**0.998**
PixelFormer [9]	CVPR'23	<u>0.090</u>	<u>0.322</u>	**0.039**	<u>0.929</u>	<u>0.991</u>	**0.998**
Ours		**0.088**	**0.319**	**0.039**	**0.931**	**0.992**	**0.998**

Bold indicates the optimal solution, underlined indicates the suboptimal solution, ↑ indicates the higher the better, ↓ indicates the lower the better.

Table 2. Comparison experiment results on the SUNRGB-D dataset.

Method	Venue	Abs Rel ↓	RMSE ↓	$\log_{10}$ ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
PackNet-SAN [30]	CVPR'21	0.166	0.494	0.071	0.757	0.943	0.984
Adabins et al. [18]	CVPR'21	0.183	0.541	0.082	0.696	0.912	0.973
Naderi et al. [8]	WACV'22	0.172	0.515	0.075	0.740	0.933	0.980
Lee et al. [27]	WACV'22	0.159	0.476	0.068	0.771	0.944	0.983
PixelFormer [9]	CVPR'23	0.144	0.441	0.062	0.802	0.962	0.990
Ours		**0.140**	**0.438**	**0.060**	**0.803**	**0.964**	**0.990**

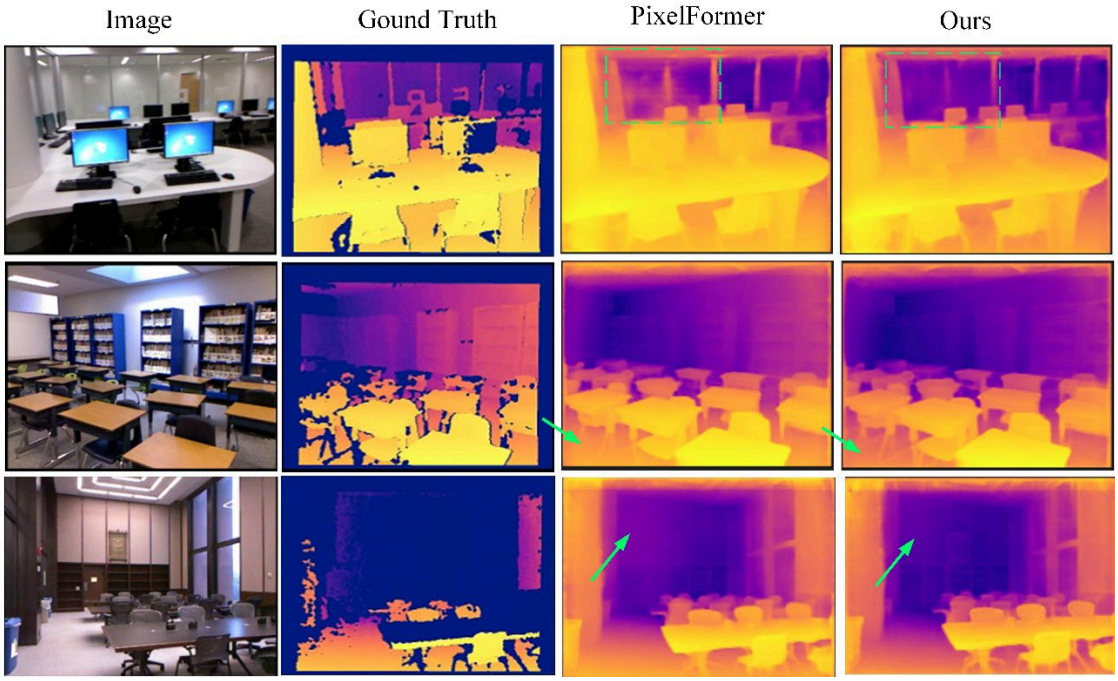


Fig. 3. Comparisons of the depth map estimation results of the PixelFormer method and the algorithm proposed in this paper on the NYU-Depth v2 dataset.

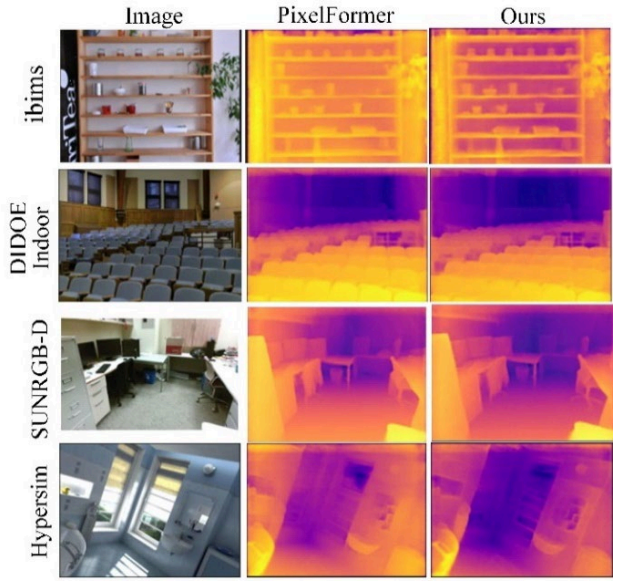


Fig. 4. Comparison of the depth map estimation results of the PixelFormer method and the algorithm proposed in this paper on four different indoor test sets.

4.4. NYU Depth v2 Dataset

To validate the effectiveness of the proposed module, we conducted stepwise ablation experiments on the NYU Depth v2 dataset, with results shown in Table 3. It can be observed that:

- Adding the TRCA module significantly reduces AbsRel and SqRel,
- δ<1.25 improves, indicating that textual semantic priors enhance the discriminative power of depth features, with particularly notable gains in texture-sparse or boundary regions;
- Incorporating the TBCP markedly reduces RMSE, demonstrating that semantic priors combined with batching strategies effectively model complex depth distributions;
- Simultaneously integrating TRCA and TBCP achieves optimal performance across all metrics, validating the synergistic complementarity of both modules in feature

Table 3. Ablation experimental results of the method proposed in this paper on the NYU v2 dataset.

Method	TDFE	TBCP	Abs Rel ↓	RMSE ↓	$\log_{10}$ ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Baseline			0.090	0.322	**0.039**	0.929	0.991	**0.998**
Baseline+TRCA	○		**0.088**	0.321	**0.039**	**0.931**	**0.992**	**0.998**
Baseline+TBCP		○	0.089	0.319	**0.039**	0.930	**0.992**	**0.998**
Ours	○	○	**0.088**	**0.318**	**0.039**	**0.931**	**0.992**	**0.998**

enhancement and depth interval modeling.

V. CONCLUSION

This paper proposes a single-image depth estimation method that integrates cross-modal semantic priors. By incorporating CLIP text information into the PixelFormer framework, we design a TRCA module and a TBCP module. These innovations effectively enhance the expressive power of depth features and improve the robustness of depth interval segmentation in complex scenes. Experiments demonstrate that our method significantly outperforms PixelFormer on NYU Depth v2 and exhibits stronger generalization capabilities across four out-of-domain benchmarks. These results indicate that cross-modal semantic priors effectively complement traditional visual supervision, offering novel insights for enhancing the generalization ability of single-image depth estimation. Moving forward, we will further explore the potential of larger-scale vision-language models for pixel-level semantic alignment, as well as the robustness of the proposed method in outdoor dynamic scenes and under extreme conditions.

ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 62366032 and Grant 62361043, in part by the Graduate Student Innovation Special Fund Project for the Jiangxi Province under Grant YC2024-S609.

REFERENCES

[1] L. Kong, S. Xie, H. Hu, L. X. Ng, B. R. Cottereau, and W. T. Ooi, "Robodepth: Robust out-of-distribution depth estimation under corruptions," *Advances in Neural Information Processing Systems*, vol. 36, pp. 21298-21342, 2023.

[2] H. Wang, S. Jia, T. Zeng, G. Zhang, and Z. Li, "Feature disentanglement in one-stage object detection," *Pattern Recognition*, vol. 145, p. 109878, 2024.

[3] S. Li, K. Huang, J. Chu, and L. Leng, "MSAF: Multi-scale adaptive filter for object tracking," *Expert Systems with Applications*, vol. 294, p. 128715, 2025.

[4] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1623-1637, 2020.

[5] A. Eftekhar, A. Sax, R. Bachmann, J. Malik, and A. Zamir, "Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3D scans," in *Proceedings of the IEEE/CVF International Conference on Computer Vision,* 2021.

[6] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, "Zoedepth: Zero-shot transfer by combining relative and metric depth," *arXiv Prep. arXiv:2302.12288*, 2023.

[7] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, "Repurposing diffusion-based image generators for monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[8] T. Naderi, A. Sadovnik, J. Hayward, and H. Qi, "Monocular depth estimation with adaptive geometric attention," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022.

[9] A. Agarwal and C. Arora, "Attention attention everywhere: Monocular depth prediction with skip attention," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, and T. Unterthiner, et al., "An image is worth 16×16 words: Transformers for image recognition at scale," *arXiv Prep. arXiv:2010.11929*, 2020.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, and S. Agarwal, et al., "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021.

[12] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[13] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale

deep network," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[14] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[15] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE International Conference on Computer Vision,* 2015.

[16] C. Wang, S. Lucey, F. Perazzi, and O. Wang, "Web stereo video supervision for depth prediction from dynamic scenes," *arXiv Prep. arXiv:1904.11112*, 2019.

[17] Z. Song, Z. Wang, B. Li, H. Zhang, R. Zhu, and L. Liu, et al., "Depthmaster: Taming diffusion models for monocular depth estimation," *arXiv Prep. arXiv: 2501.02576*, 2025.

[18] S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[19] W. Yin, Y. Liu, and C. Shen, "Enforcing geometric constraints of virtual normal for depth prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[20] X. Qi, R. Liao, Z. Liu, R. Urtasun, and J. Jia, "Geonet: Geometric neural network for joint depth and surface normal estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[21] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[22] S. Kim, J. Kang, D. Kim, and S. Lee, "CLIP can understand depth," *Pattern Recognition*, p. 112475, 2025.

[23] A. Chatterjee, T. Gokhale, C. Baral, and Y. Yang, "On the robustness of language guidance for low-level vision tasks: Findings from depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[24] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, and Z. Zhang, et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[25] V. Patil, C. Sakaridis, A. Liniger, and L. Van Gool, "P3depth: Monocular depth estimation with a piecewise planarity prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[26] C. Couprie, C. Farabet, L. Najman, and Y. LeCun, "Indoor segmentation and support inference from rgbd images," in *European Conference on Computer Vision*, 2012.

[27] J. H. Lee, M. K. Han, D. W. Ko, and I. H. Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation," *arXiv Prep. arXiv: 1907.10326*, 2019.

[28] L. Huynh, P. Nguyen-Ha, J. Matas, E. Rahtu, and J. Heikkila, "Guiding monocular depth estimation using depth-attention volume," in *European Conference on Computer Vision*, 2020.

[29] G. Yang, H. Tang, M. Ding, N. Sebe, and E. Ricci, "Transformer-based attention networks for continuous pixel-wise prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[30] V. Guizilini, R. Ambrus, W. Burgard, and A. Gaidon, "Sparse auxiliary networks for unified monocular depth prediction and completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Journal of Multimedia Information System VOL. 13, NO. 2, June 2026 (pp. 43-52): ISSN 2383-7632 (Online)
https://doi.org/10.33851/JMIS.2026.13.2.43

AUTHORS

Sheng Yu received the B.S. degree in Mechanical Engineering from Nanchang Jiaotong University in 2023. Currently pursuing a M.S. degree at the Computer Vision Research Institute of Nanchang Hang Kong University. Research focuses on monocular depth estimation.

Jun Miao received the M.S. degree in Mechanical Design and Theory from North-western Poly technical University, Xi'an, China, in 2008, and the Ph.D. degree from Nanchang University, Nanchang, China, in 2015. He was a Visiting Scholar at the University of California, Merced, Merced, CA, USA. He is currently a Researcher at the Key Laboratory of Jiangxi Province for Image Processing and Pattern Recognition, and a Professor in the School of Aeronautical Manufacturing and Mechanical Engineering, Nanchang Hangkong University. His research interests include computer vision, 3D reconstruction, and predictive modeling of machining parameters.

Jianqiao Wang received the B.S. degree in software engineering from Jiangxi University of Finance and Economics in 2023. He is currently working toward a M.S. degree at the Computer Vision Research Institute of Nanchang Hangkong University. His research interests include 3D reconstruction and new view synthesis.