Automating Prompt Generation for Training-Free Object Segmentation in PaintSeg

Pranav Mishra
Department of Computer Science
University of Illinois Chicago
Chicago, IL
pmishr23@uic.edu

Annesh Potnis
Department of Computer Science
University of Illinois Chicago
Chicago, IL
apotni2@uic.edu

Aditya Pimpley
Department of Computer Science
University of Illinois Chicago
Chicago, IL
apimp@uic.edu

Abstract

Object segmentation in visual computing has traditionally relied on manual or semi-supervised methods that require extensive training and human intervention for prompt specification. This paper introduces an innovative autoprompting system designed to automate the input generation for PaintSeg, a training-free and unseen object segmentation model. Our methodology leverages k-means clustering for color-based segmentation and employs the Dense Prediction Transformer (DPT) model to extract depth maps, creating precise binary and bounding box masks without manual input. Conducted experiments on the DUTS dataset demonstrated that our autoprompting approach achieves Intersection Over Union (IOU) scores between 45 and 55 percent, with an enhancement of up to 60 percent through a hybrid prompting strategy that intelligently combines mask types based on their spatial characteristics. This work not only streamlines the segmentation process but also opens new avenues for further automation in image processing tasks.

1 Introduction

A key job in visual computing is object segmentation, which is usually hampered by its reliance on manual or semi-supervised approaches that require human intervention for swift specification and substantial instruction. This reliance not only reduces efficiency but also inhibits segmentation models' capacity to adapt to many different and unknown objects. Automating this procedure appears to be a significant development since it could simplify processes and increase the range of applications for segmentation technology.

Providing training-free segmentation of unseen objects, the PaintSeg model [13] is a breakthrough solution to this problem. This novel model greatly diminishes the reliance on large pre-trained datasets, enabling a wider range of applications. Our work leverages PaintSeg's capabilities by creating a reliable auto-prompting system that generates input masks automatically.

Our research addresses this bottleneck by automating the generation of input masks for PaintSeg. We utilize two parallel approaches: K-means clustering for color-based segmentation and the Dense Prediction Transformer (DPT) model for depth-based segmentation [15]. While K-means provides

a robust, training-free method well-suited to the philosophy of PaintSeg, it has limitations, particularly in complex scenes with subtle color distinctions or variable lighting. To complement this, we integrate DPT, a trained model that excels in capturing depth differentials across objects, providing a richer context for segmentation tasks.

Following are our paper's key contributions:

- **Introduction of an Automated Prompting Strategy:** We introduce a fully automated approach that makes use of depth perception and/or sophisticated clustering techniques to produce high-quality PaintSeg masks, improving the model's efficiency and scope of use.
- **Hybrid Approach for Mask Generation:** To solve the various issues encountered in object segmentation, our methodology combines the advanced capabilities of the DPT model [15] with the strengths of unsupervised K-means clustering in a distinctive manner.
- **Empirical Validation and Improved Performance:** Extensive experiments on the DUTS dataset [18] show that our method outperforms conventional approaches in terms of performance and also improves the process of choosing between various mask types by employing a strategic overlap thresholding technique that greatly increases the accuracy of segmentation.

2 Related Work

2.1 Advancements in Generative and Autonomous Segmentation

The evolution of image segmentation has been significantly propelled by developments in deep learning, with a notable trend toward the integration of generative models and segmentation techniques. Applications of denoising diffusion probabilistic models (DDPM) for tasks such as image inpainting and text-to-image synthesis have shown remarkable results [17, 5]. These applications highlight how the integration of image generation with segmentation can enhance the precision of generative models and improve the contextual understanding within segmentation tasks [12, 7].

Further advancements have been observed with the use of generative adversarial networks (GANs) in unsupervised segmentation methods, demonstrating the capability for autonomous image segmentation [8, 24]. This approach enriches the generative models by enabling them to learn and adapt autonomously from the segmentation processes [22, 10]. Additionally, self-supervised methods employing Vision Transformers (ViTs) have effectively utilized these models to segment objects without manual annotation, significantly improving segmentation accuracy [2, 19].

2.2 Depth-Based and Prompt-Guided Segmentation Techniques

Depth-based segmentation strategies have become crucial, particularly those that utilize both synthetic RGB and depth data to delineate object boundaries more effectively [20]. These strategies underscore the importance of depth cues in the Dense Prediction Transformer (DPT) model [15] used in our project, which manages complex segmentation challenges by integrating these depth data.

In the realm of prompt-guided segmentation, leveraging user-defined inputs like masks and points has become indispensable. This method is especially prevalent in semi-supervised video object segmentation (VOS) and interactive segmentation (IS), where user interactions refine the segmentation outputs [1, 21, 9]. Techniques such as these highlight the dynamic adaptability of segmentation models to user input and real-time data [14, 11].

2.3 Our Approach

Addressing the limitations of purely color-based segmentation methods such as K-means, our approach integrates both classical and adaptive K-means clustering techniques. Research has shown that these methods can be enhanced significantly by pre-processing and adaptive thresholding [4, 23], thus improving the overall performance of segmentation tasks. To put it briefly, our approach synthesizes these several lines of research by fusing the depth-aware segmentation made possible by DPT with the powerful, unsupervised capabilities of K-means. This hybrid technique

pushes the limits of what automated systems can accomplish in practical applications, while still keeping up with and improving upon current trends in image segmentation.

3 Methodology

In our research, we focus on two main techniques to generate input masks for PaintSeg, leveraging both K-means clustering and DPT models. This approach is designed to produce high-quality binary masks that are subsequently refined into either coarse masks or bounding box masks, depending on their clarity and effectiveness in representing the segmented objects. .

3.1 Kmeans clustering for segmentation

The first step in our methodology involves using K-means clustering to simplify the color space of the images. By reducing the images to a smaller set of representative colors (3, 6, or 10 colors based on the color index of the image), we enhance the effectiveness of edge detection. This reduction in color complexity helps in clearly delineating the object edges, making the segmentation task less noisy and more precise.

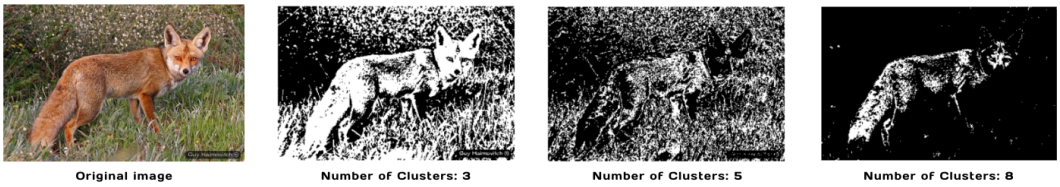

Figure 1: **Effect of Cluster Variation in K-means on Segmentation Quality**

In this process, we measure the colorfulness of each image to determine the appropriate number of clusters (k). Images with lower colorfulness metrics use fewer clusters, as they inherently have less color variability, whereas more colorful images require a higher number of clusters to accurately represent their color diversity. This step is crucial as it directly impacts the subsequent edge detection and object identification phases.

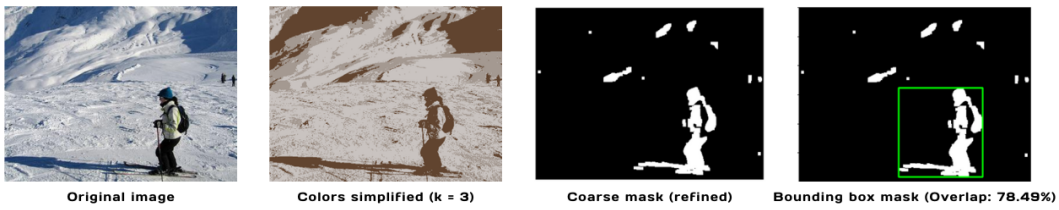

Figure 2: **K-means Clustering for Segmentation:** Sequential Steps of clustering techniques for coarse mask generation as well as refined bounding box masks.

After applying K-means clustering, we extract significant color blobs within the images. By focusing on the largest blobs based on area and ranking them, we can selectively refine our object representations. This selective focus helps in preparing the masks for further processing, where we enhance the object's boundary definitions through morphological operations and connected components analysis.

3.2 DPT depth maps for segmentation

Separately, we employ the DPT model [15] to generate depth maps for each image. These monochrome outputs indicate the depth information, where brighter areas represent closer to the camera and darker areas signify the background. Using this depth information, we perform a range thresholding to isolate the most prominent objects based on depth. This process simplifies the seg-

mentation by focusing on significant depth discrepancies, which typically correspond to the main objects in the scene.

Figure 3: **Depth-Based Segmentation Using the DPT Model:** Sequential Steps of DPT Model Segmentation: From depth map generation to refined bounding box mask.

The binary masks generated from the DPT outputs are further refined using edge detection techniques. By overlaying the detected edges onto the binary masks, we can make nuanced adjustments to the mask boundaries. This overlay helps in removing extraneous pixels that do not contribute to the core object shape and in including essential interior details that may be surrounded by edges.

3.3 Choosing Between Coarse and Bounding Box Masks

Both the K-means and DPT methodologies produce two types of masks: coarse masks and bounding box masks. The coarse masks are derived from the blob analysis and edge refinement processes, focusing on the main object areas as identified through color and depth analysis. Bounding box masks, on the other hand, are generated by enclosing the largest connected components and their neighboring elements within a calculated boundary, adjusted to avoid edge overlaps and ensure comprehensive coverage of the object.

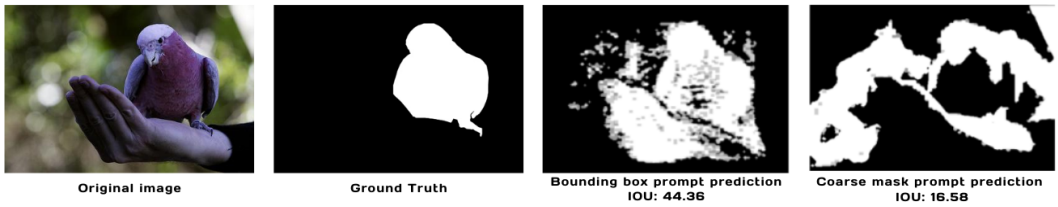

Figure 4: **Comparative Analysis of Segmentation Techniques:** Performance comparison of bounding box and coarse mask prompts against the ground truth for PaintSeg segmentation, highlighting how bounding box prompts can outperform coarse mask prompts at times.

To determine the most effective mask type for each image, we calculate the overlap percentage between the coarse and bounding box masks. This metric helps in deciding whether the bounding box or the coarse mask provides a better representation of the object based on how much of the object each type encompasses relative to the other.

3.3.1 Final Selection and Mask Differentiation

In our methodology, the final mask selection for input into PaintSeg results in two distinct sets of masks derived from K-means clustering and DPT models, respectively. This bifurcation is essential as it allows us to harness specific advantages from each technique, adapting to the unique characteristics and demands of each image in the dataset.

For each image, the K-means and DPT approaches independently determine whether a coarse or bounding box mask better captures the object's details. For instance, consider a dataset with three images. The K-means process might yield a selection of [coarse, coarse, bounding box], depending on which mask type more accurately encompasses the salient features of the object in each respective image. Conversely, the DPT-derived masks for the same images could differ, possibly resulting in [coarse, bounding box, coarse] configurations.

Crucially, the masks produced by K-means and DPT are not only different in terms of the mask type (coarse or bounding box) but also in their internal characteristics and the specifics of the segmentation. K-means masks are primarily influenced by color segmentation and blob analysis, which may lead to a certain interpretation of object boundaries and core areas. On the other hand, DPT masks, derived from depth cues and refined through edge detection, may highlight different aspects of the objects based on depth variations and contour dynamics.

3.3.2 Why the distinction

This dual-pathway approach to mask generation underscores the fundamental separation between the methods used. While the DPT model relies on a trained algorithm to produce depth-based segmentation outputs, K-means clustering operates as a completely training-free technique, aligning more closely with the core philosophy of PaintSeg. By distinctly segregating the masks generated from each method, we ensure that the inherently different qualities of both approaches are preserved. The K-means method enhances the dataset with purely unsupervised, training-free segmentation masks, providing a direct comparison against the model-dependent outputs of DPT. This clear distinction allows us to critically evaluate the benefits and limitations of using trained versus untrained methodologies in object segmentation. In subsequent sections, we will delve into the comparative analysis of these approaches, highlighting how each contributes to tackling diverse segmentation challenges within PaintSeg, thereby facilitating robust, training-free object segmentation in varied scenarios.

4 Implementation

4.1 Tools and Computational Environment

Several computing environments were utilized to manage various stages of our process efficiently. The Dense Prediction Transformer (DPT) models, crucial for producing monodepth maps, were operated on Google Collab's GPUs [16]. These depth maps are fundamental for generating one set of input masks for the PaintSeg model.

For initial segmentation and mask generation using K-means clustering, along with other post-processing tasks, we used a Python environment on our personal computing systems. This setup allowed us to efficiently handle the computational demands of color simplification and blob detection tasks, which are less computationally intensive than deep learning model operations.

Once the two distinct sets of masks from the DPT model and K-means processing were obtained, we used them as inputs for the PaintSeg model [6]. This model required a specific Linux-based environment for optimal performance, as outlined in its GitHub repository. Due to hardware constraints, specifically our access to an Nvidia 4060 GPU, we were limited to only five iterations of the IO steps/diffusion steps prescribed by PaintSeg, despite the model supporting up to 50 iterations, thus limiting the depth of our analysis.

4.2 Dataset Utilization

We employed the DUTS dataset, designed for saliency detection, which includes a large-scale dataset with explicit training and test evaluation protocols [3]. This dataset consists of 10,553 training images and 5,019 test images, sourced from ImageNet DET training/val sets and the SUN dataset for test images, providing a challenging set of scenarios for saliency detection.

Due to computational and resource constraints, our experiments were selectively conducted on subsets of the DUTS dataset. We randomly sampled several sets, each containing hundreds of images, to process through our dual-methodology pipeline. This approach allowed us to manage the computational load effectively while still leveraging the robust capabilities of both K-means clustering and DPT-derived masks as inputs into the PaintSeg model. This sampling strategy ensured that our findings remained statistically relevant and reflective of the dataset's overall characteristics, despite not utilizing the entire dataset in our experiments.

4.3 Findings

We saw notable improvements in the segmentation results in both pathways, even with the hardware limitations that limited us to five iterations of PaintSeg's capabilities. These initial findings imply that the dual-pathway strategy to generate pre-segmentation masks can significantly improve the performance of training-free segmentation models such as PaintSeg, even with limited iteration counts.

Table 1: Metrics - IOU Scores across different segmentation models and techniques.

MODEL	KMEANS	KMEANS + PAINTSEG	DPT	DPT + PAINTSEG
Coarse Mask: IOU Score	44.34	51.89	49.63	57.92
Bounding box: IOU score	-	41.29	-	49.33
Selective Prompting : IOU score	-	57.54	-	63.61
(DPT + Kmeans) + PAINTSEG	Best of four: Selective prompting		70.98	
	Overlay Masks:		72.48	

5 Results

5.1 Kmeans Prompting Techniques

Initially, we employed K-means clustering to generate coarse masks for our images, and without any iterations of PaintSeg's diffusion steps, these masks yielded an IOU (Intersection Over Union) score of 44.34%. The IOU score, a standard metric for evaluating object segmentation, measures the overlap between the predicted mask and the ground truth mask as a percentage, highlighting the accuracy of our segmentation.

- **Coarse masks vs bounding box masks:** When these coarse masks were further processed through the PaintSeg model for five iterations, the resulting masks exhibited a significant improvement, with an IOU of 51.89%. We also experimented with bounding box inputs derived from the initial K-means masks; however, these achieved a lower IOU of 41.29% after five iterations in PaintSeg.
- **Selective Prompting (via Overlap Thresholding)** Upon evaluating the overlap between the initial coarse masks and bounding boxes, we established a threshold value of approximately 70% overlap. Images with an overlap exceeding this threshold were better suited to bounding box prompting, and vice versa, leading to a further refined IOU of 57.54% when using a selective prompting approach in PaintSeg.

5.2 DPT Prompting Techniques

For the DPT-derived masks, the coarse masks established through range thresholding alone provided an IOU of 49.63%. These were then input into PaintSeg, enhancing the IOU to 57.92% after five iterations. Similar to the K-means approach, bounding boxes generated from the initial coarse masks yielded an IOU of 49.33% after processing with PaintSeg.

Implementing the same overlap thresholding technique between the initial coarse masks and bounding box masks, with adjustments based on a 70% threshold, we improved the IOU to 63.61% after five iterations of PaintSeg, demonstrating the effectiveness of selecting prompts based on their overlap percentages.

5.3 Combined Techniques

- **Best of four:** In our combined approach, we incorporated masks from both K-means and DPT methodologies. Initially, we applied the overlap thresholding between all four types of masks (two sets of coarse and two sets of bounding box masks), selecting the mask with

the best overlap percentage for each image. This meticulous selection process elevated the IOU to 70.98%.

- **Overlay Mask sets:** We also explored creating overlaid masks by utilizing the common pixels between the coarse masks of both K-means and DPT methods, and similarly for the bounding boxes. This new set of masks, following our established overlap thresholding criteria, resulted in an even higher IOU of 72.48%. This approach effectively leveraged the strengths of both K-means and DPT masks, compensating for the individual limitations of each method.

Figure 5: **Overlay Masks:** Comparison of segmentation masks generated using K-means, DPT, and their overlay.

6 Discussion

The results indicate that DPT masks generally outperform K-means masks in both coarse and bounding box configurations, with initial IOUs of 49.63% for coarse and 49.33% for bounding boxes compared to K-means' scores of 44.34% and 41.29%, respectively. The combination of K-means coarse and bounding box masks using overlap thresholding significantly improved outcomes, suggesting that while K-means masks are less effective in scenarios with high contrast or complex lighting conditions, they can still be optimized through strategic use of bounding boxes.

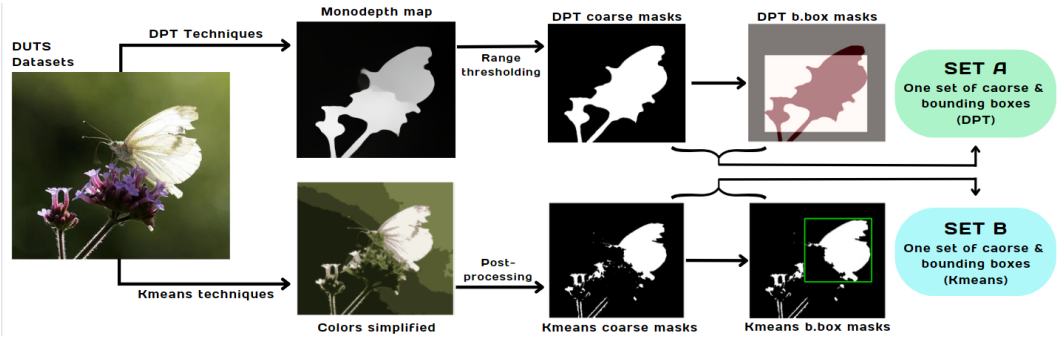

Figure 6: **Flowchart illustrating the segmentation process using DPT and K-means techniques:** This flowchart delineates the distinct steps involved in generating segmentation masks using two different techniques: DPT (Dense Prediction Transformer) and K-means clustering. The image here is an example from the DUTS dataset used to describe how each image is handled.

In contrast, the DPT technique did not show a substantial improvement with the overlap thresholding approach, likely due to the nature of depth maps, which often capture multiple objects at the same depth, leading to less accurate segmentations when the ground truth focuses on more specifically defined objects. However, it still pushed the IOU score to 63.61% from 57.92%.

By combining both K-means and DPT techniques, we leveraged the complementary strengths of each method. Where K-means faltered due to lighting and contrast, DPT's depth cues provided more robust segmentation, and vice versa for scenarios involving multiple objects in the same depth. This

synergy allowed us to push the IOU to 72.48%, illustrating that a hybrid approach can significantly enhance segmentation accuracy, particularly in challenging imaging conditions.

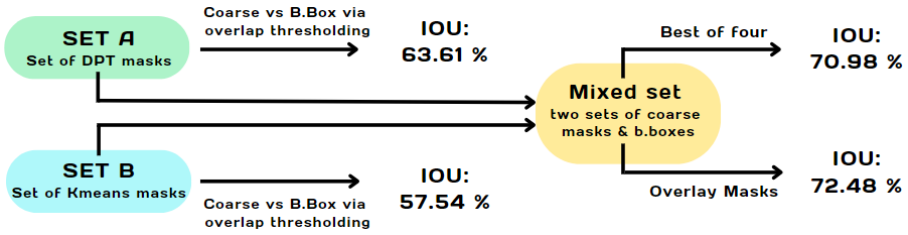

Figure 7: **Combination and Evaluation of Mask Sets:** Visualization of the process and performance evaluation for combining K-means and DPT mask sets. SET A (DPT masks) and SET B (K-means masks) are initially processed through an overlap thresholding approach to optimize mask selection. The results are then combined in two strategies: Best of Four and Overlay Masks, which ultimately enhance the Intersection Over Union (IOU) scores, demonstrating the effectiveness of hybrid segmentation approaches in complex imaging environments

7 Conclusion

In our study, we explored the integration of two distinct segmentation methods—K-means clustering and DPT models—to generate input prompt masks for PaintSeg, a training-free segmentation model. Each method offered unique advantages, with K-means providing a purely unsupervised approach, aligning closely with PaintSeg's philosophy, and DPT offering depth-based insights that enhanced segmentation quality. Our methodology proved effective, demonstrating that high-quality binary masks, refined into either coarse or bounding box masks based on their performance, can significantly enhance segmentation outcomes. This study underscores the potential of hybrid approaches in tackling complex segmentation tasks and sets a foundation for further exploration into effective combinations of trained and untrained methodologies in image segmentation.

Limitations

While the combined approach using both K-means and DPT masks yielded an IOU of up to 72.48%, it is important to note that this is not purely a training-free approach due to the trained nature of the DPT model. However, using purely K-means techniques, we achieved an IOU of over 57%, which does align with the training-free, unsupervised segmentation ethos of the PaintSeg model.

References

[1] Ali Athar, Alexander Hermans, Jonathon Luiten, Deva Ramanan, and Bastian Leibe. Tarvis: A unified approach for target-based video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18738–18748, 2023.

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, October 2021.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[4] Nameirakpam Dhanachandra, Khumanthem Manglem, and Yambem Jina Chanu. Image segmentation using k -means clustering algorithm and subtractive clustering algorithm. *Procedia Computer Science*, 54:764–771, 2015. Eleventh International Conference on Communication Networks, ICCN 2015, August 21-23, 2015, Bangalore, India Eleventh International Conference on Data Mining and Warehousing, ICDMW 2015, August 21-23, 2015, Bangalore, India Eleventh International Conference on Image and Signal Processing, ICISP 2015, August 21-23, 2015, Bangalore, India.

[5] Qiaole Dong, Chenjie Cao, and Yanwei Fu. Incremental transformer structure enhanced image inpainting with masking positional encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11358–11368, June 2022.

[6] Xiang Li et al. PaintSeg: Training-free segmentation via painting. https://github.com/lxa9867/PaintSeg, 2023. Accessed: 2024-04-30.

[7] Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Trans. Graph.*, 41(4), jul 2022.

[8] Federico Galatolo, Mario Cimino, and Gigliola Vaglini. Generating images from caption and vice versa via clip-guided generative latent space search. In *Proceedings of the International Conference on Image Processing and Vision Engineering*. SCITEPRESS - Science and Technology Publications, 2021.

[9] Yuying Hao, Yi Liu, Zewu Wu, Lin Han, Yizhou Chen, Guowei Chen, Lutao Chu, Shiyu Tang, Zhiliang Yu, Zeyu Chen, et al. Edgeflow: Achieving practical interactive segmentation with edge-guided flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1551–1560, 2021.

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[12] Xiang Li, Yinpeng Chen, Chung-Ching Lin, Hao Chen, Kai Hu, Rita Singh, Bhiksha Raj, Lijuan Wang, and Zicheng Liu. Completing visual objects via bridging generation and segmentation, 2024.

[13] Xiang Li, Chung-Ching Lin, Yinpeng Chen, Zicheng Liu, Jinglu Wang, Rita Singh, and Bhiksha Raj. Paintseg: Painting pixels for training-free segmentation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 35–56. Curran Associates, Inc., 2023.

[14] Yongqing Liang, Xin Li, Navid Jafari, and Jim Chen. Video object segmentation with adaptive feature bank and uncertain-region refinement. *Advances in Neural Information Processing Systems*, 33:3430–3441, 2020.

[15] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, October 2021.

[16] René Ranftl et al. Dense prediction transformers. https://github.com/isl-org/DPT, 2021. Accessed: 2024-04-30.

[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[18] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, 2017.

[19] Xudong Wang, Rohit Girdhar, Stella X. Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3124–3134, June 2023.

[20] Christopher Xie, Yu Xiang, Arsalan Mousavian, and Dieter Fox. Unseen object instance segmentation for robotic environments. *IEEE Transactions on Robotics*, 37(5):1343–1359, 2021.

[21] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Shengping Zhang, and Wenxiu Sun. Efficient regional memory network for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1286–1295, 2021.

[22] 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 (ICCV)*, pages 3836–3847, October 2023.

[23] Xin Zheng, Qinyi Lei, Run Yao, Yifei Gong, and Qian Yin. Image segmentation based on adaptive k-means algorithm. *EURASIP Journal on Image and Video Processing*, 2018(1):68, 2018.

[24] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17907–17917, June 2022.