Cross-modal Consistent Augmentation bridging 2D–3D Transformations

Donguk Kim
Department of Automotive Engineering
Hanyang University
Seoul, Republic of Korea
donguk0513@hanyang.ac.kr

Soonmin Hwang
Department of Automotive Engineering
Hanyang University
Seoul, Republic of Korea
soonminh@hanyang.ac.kr

***Abstract*—LiDAR–camera fusion models are essential for robust 3D scene understanding in autonomous driving. However, existing multi-modal augmentation methods are typically developed independently for the LiDAR and image domains, which often leads to misalignment between 3D geometric transformations and corresponding 2D visual information. In this paper, we present a geometry-aware framework bridging 2D-3D transformations. With a diffusion inpainting model conditioned on transformed 3D bounding boxes and object appearance, augmented images are generated to preserve LiDAR object geometry and align with the surrounding scene context. Experiments on the nuScenes dataset with BEVFusion demonstrate that our augmentation improves geometric alignment and yields consistent gains in 3D detection performance over both object sampling and object paste baselines. Our findings highlight the importance of geometry-aware, multi-modal augmentation for advancing LiDAR-Camera fusion models.**

***Index Terms*—3D object detection, data augmentation, diffusion, autonomous driving**

I. Introduction

Multi-modal information from LiDAR and camera sensors is essential for accurate 3D scene understanding in Autonomous Driving (AD). LiDAR provides accurate 3D spatial geometry of the scene, whereas cameras offer complementary 2D appearance and semantic texture such as color, texture, lighting, and contextual cues. These complementary properties make LiDAR–camera fusion a critical component of modern autonomous perception systems.

However, most existing data augmentation pipelines for multi-modal fusion still treat the two modalities independently. In the LiDAR branch, 3D augmentations such as flipping, rotation, scaling, object sampling, and scene-level mixing (e.g., PolarMix [1], LaserMix [2]) are commonly applied. In contrast, the camera branch relies on 2D operations such as flipping, rotation, cut-and-paste, and mosaic [3]. Beyond simple global transformations, LiDAR-specific 3D operations do not have corresponding camera-side transformations. As a result, when LiDAR objects undergo unshared rotations or resampling, their projected shapes and locations no longer match the static RGB images, which degrades cross-modal feature alignment.

† Corresponding author.

This misalignment issue becomes more pronounced in widely used architectures such as BEVFusion [4]. It is commonly used two-stage training strategy first pretrains a LiDAR-only detector with strong 3D augmentations and then introduces camera inputs, which tends to bias learning toward the LiDAR branch. As observed in [5], image features are often under-exploited, making it difficult for the model to develop a well-balanced multi-modal representation.

Cross-modal misalignment is often addressed using object-paste (GT-Paste [6]), which pastes corresponding 2D object patches into images based on projected 3D locations. Although the same object can be observed in both modalities, depth ordering, occlusion, and scene-level appearance are not explicitly considered, which often results in visible artifacts. PointAugmenting [7] considers depth and occlusion in both LiDAR and camera views, which improves geometric alignment. However, since it is still based on cut-and-paste in the image domain, mismatches in illumination, color, and surrounding context are not fully resolved.

To overcome these limitations, we introduce a diffusion-based multi-modal augmentation framework that keeps LiDAR and camera views geometrically aligned. We first apply 3D transformations to LiDAR objects, project the resulting 3D bounding boxes onto each camera image, and then perform Stable Diffusion [8]-based inpainting conditioned on the projected geometry to generate context-compatible object appearance. While the object's position and scale remain fixed, its texture, lighting, and seam transitions are blended with the target scene, which mitigates typical cut-and-paste artifacts. On BEVFusion [4], our method yields consistent gains over object sampling and object-paste baselines by strengthening cross-modal alignment.

II. Related Work

Heterogeneous Augmentation for Multi-modal Fusion. Existing augmentation techniques widely used in LiDAR–camera fusion models remain largely confined to modality-specific approaches. For LiDAR, basic 3D point-based transformations such as flipping, rotation, and scaling are primarily employed, alongside techniques like instance sampling or scene-level mixing (e.g., PolarMix [1], LaserMix [2]). In contrast, the camera domain employs 2D image transformation-based

augmentations like flipping, resizing, color jittering, cut-and-paste, and mosaic [3], which are applied independently of 3D geometric transformations. Due to the inherent heterogeneity between the two modalities, consistency is generally limited to simple global transformations such as scene rotation or flipping. This limitation restricts the range of augmentation strategies that can be directly exploited by multi-modal fusion models. To mitigate these issues, PointAugmenting [7] introduces an occlusion-aware cross-modal augmentation framework that leverages depth information to explicitly account for occlusion during object insertion in images. While this approach effectively preserves geometric consistency, it does not adequately capture photometric properties or contextual coherence.

Diffusion. Diffusion models (DDPM [9], LDM [8]) stabilize high-resolution synthesis and inpainting compared to GAN [10] and VAE [11] approaches but largely operate in 2D with text or label conditioning, offering limited explicit 3D control. Recent 3D-aware efforts (e.g., MagicDrive [12] with text, HD maps, BEV features, and 3D boxes) primarily target scene generation or simulation rather than augmentation for LiDAR–camera fusion. Closer to augmentation, MObI [13] conditions a Paint-by-example [14] diffusion model on range images and 3D boxes to refine image realism, but it functions as an image-level inpainting stage and does not integrate with common LiDAR object sampling strategy. Neural Assets [15] enables controllable multi-object 2D synthesis by disentangling appearance and pose, but it is computationally heavy and has not been evaluated for LiDAR–camera fusion or downstream detection. In contrast, we directly leverage 3D-aware diffusion conditioning to build geometry-consistent, cross-modal augmentations that plug into standard LiDAR object sampling and scene mixing, delivering both 3D alignment and photometric and contextual realism for training-time fusion.

III. Method

This section describes cross-modal geometry-consistent augmentation framework in detail. The framework consists of three components: (1) procedures for baseline object sampling and object pasting, (2) a geometry-aware diffusion inpainting model conditioned on 3D boxes, and (3) a geometry-consistent cross-modal augmentation framework.

A. Object Sampling and Object-Paste

We first introduce the object sampling and object-paste baselines used for comparison. Given an original scene (e.g. Figure 1-(a)), we take the point cloud of source object $\mathcal{P}_{\text{obj}}$ and its 3D bounding box $B_{\text{obj}} = (c, d, \theta)$, where $c \in \mathbb{R}^3$ is the object center, $d = (w, h, l)$ the object size, and θ the yaw angle.

Object sampling. As shown in Figure 1-(b), the extracted object points are directly inserted into a target scene:

$$\tilde{\mathcal{P}}_{\text{obj}} = T_{\text{src}\to\text{tgt}} \cdot \mathcal{P}_{\text{obj}}, \qquad \tilde{B}_{\text{obj}} = T_{\text{src}\to\text{tgt}} \cdot B_{\text{obj}}, \quad (1)$$

where $T_{\text{src}\to\text{tgt}}$ is the rigid transformation applied to align the source object with the target LiDAR coordinate system. While this procedure preserves geometric consistency in the LiDAR domain, the camera image remains unchanged, resulting in cross-modal misalignment.

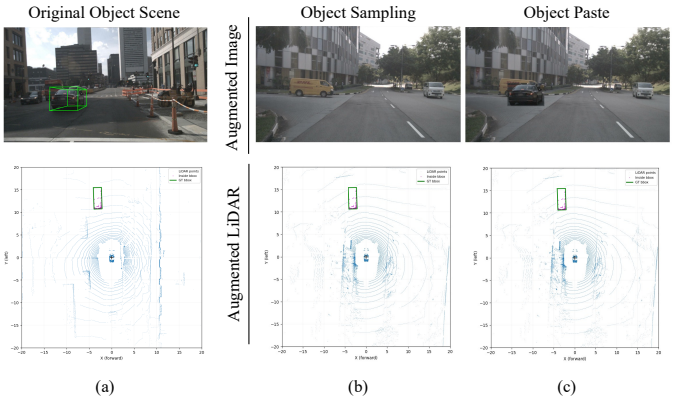

Fig. 1: (a) Original scene with the source object. (b) Object sampling: The point cloud and 3D bounding box of the source object are transformed and inserted into the target scene to augment the LiDAR point cloud without modifying the image. (c) Object-paste: The sampled object is inserted into both modalities, preserving geometric consistency but often leading to visual artifacts in appearance and context.

Object-paste. To paste the object into the image domain, we compute the eight transformed 3D bounding box corners of $\tilde{B}_{\text{obj}}$ and project them into the image:

$$u = \{u_i\}_{i=1}^{8} \in \mathbb{R}^{8\times 2}. \quad (2)$$

A 2D bounding box is obtained as

$$\tilde{B}_{\text{2D}} = (\min_i u_i^x, \min_i u_i^y, \max_i u_i^x, \max_i u_i^y), \quad (3)$$

and the corresponding patch is cropped and blended into the target image (Figure 1-(c)). Although this produces a shared object in both modalities, naive cut-and-paste often introduces artifacts such as color discontinuities and background mismatch.

B. 3D Geometry-Aware Diffusion Inpainting

As shown in Figure 2, our 3D geometry-aware diffusion inpainting conditions on projected 3D boxes to preserve LiDAR–camera alignment. We build upon the MObI [13] architecture but simplify the modality pathways by removing the range-view branch and operating solely on the camera image. To enforce geometric consistency between LiDAR and image domain, the model receives three conditioning inputs: (1) a reference image providing object appearance, (2) a LiDAR-derived 3D bounding box describing object geometry, and (3) a target image where the object region is masked. Each input is encoded into latent conditioning tokens and injected into the U-Net [16] denoiser via cross-attention.

Inpainting formulation. We formulate inpainting by masking out the object region defined by the projected 3D box. Following Stable Diffusion v2.1, the model does not observe the full

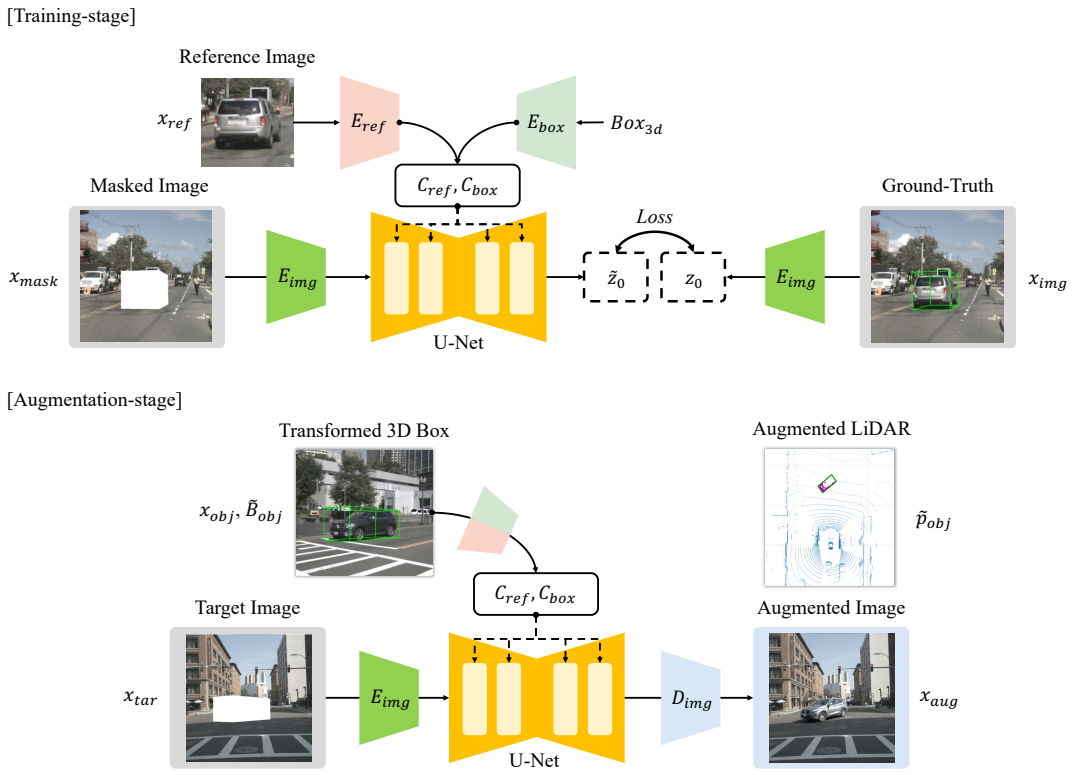

Fig. 2: Overview of the proposed cross-modal geometry-consistent augmentation framework.

target image; instead, the masked image is constructed using a binary mask $m \in \{0,1\}^{D\times D}$:

$$x_{\text{mask}} = x_{\text{img}} \odot (1-m).1 \tag{4}$$

A pretrained VAE encoder E_{img} maps the masked input to the latent space:

$$z_0 = E_{\text{img}}(x_{\text{mask}}) \in \mathbb{R}^{d\times d\times c}, \tag{5}$$

where $d = D/s$ is determined by the VAE down-sampling factor s.

During the forward process of diffusion, noise is added according to

$$z_t = \sqrt{\alpha_t}\, z_0 + \sqrt{1-\alpha_t}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{6}$$

and the U-Net denoiser ϵ_θ predicts the noise under the combined conditioning

$$C = \{C_{\text{ref}}, C_{\text{box}}\}. \tag{7}$$

The model is trained using the standard diffusion noise-prediction objective function:

$$\mathcal{L} = \mathbb{E}_{z_0,\, t,\, \epsilon}\left[\|\epsilon - \epsilon_\theta(z_t, C, t)\|^2\right], \tag{8}$$

which reconstructs the masked region in a manner consistent with the reference appearance and the 3D geometric constraints.

Reference image encoding. The reference image x_{ref} is obtained by cropping the minimal 2D region covering the projected 3D bounding box. We encode x_{ref} using a pretrained CLIP [17] image encoder E_{ref}, and pass the resulting embedding through a trainable MLP layer to obtain the appearance token:

$$C_{\text{ref}} = MLP_{\text{ref}}\big(E_{\text{ref}}(x_{\text{ref}})\big), \tag{9}$$

which preserves object-level cues such as color, texture, and fine-grained shape.

3D bounding box encoding. The 3D bounding box is projected onto the image using calibrated extrinsics and intrinsics, yielding an eight points representation $Box_{3d} \in \mathbb{R}^{8\times 3}$ encoding (x, y) image coordinates. Following the 3D box encoding strategy of MagicDrive [12], we apply Fourier positional encoding followed by an MLP layer:

$$C_{\text{box}} = E_{\text{box}}(Box_{3d}) = MLP_{\text{box}}(Fourier(Box_{3d})) \tag{10}$$

This geometry conditioned token provides the U-Net with implicit spatial cues that guide it to synthesize image content aligned with the transformed object.

C. Geometry-Consistent Cross-Modal Augmentation

We now use the trained diffusion model to generate geometry-consistent augmentation pairs for LiDAR-camera fusion. Although the conventional object sampling pipeline is capable of inserting multiple objects into a target scene, current diffusion models struggle to perform stable multi-object inpainting, and multi-object generation frameworks

Fig. 3: Comparison of object-paste and diffusion-based augmented image.

such as Neural Assets [15] incur prohibitive computational costs. Therefore, our augmentation procedure operates at the single-object level while still following the standard object sampling setup.

The process begins with the object sampling step described in Section III-A. The source object is transformed into the target LiDAR frame, producing a transformed 3D box $\tilde{B}_{\text{obj}}$ and point cloud $\tilde{\mathcal{P}}_{\text{obj}}$, where the transformed 3D box is used to condition the diffusion model on object geometry. Using the projected 3D box, a patch is cropped from the source image to extract an appearance token, while the transformed box $\tilde{B}_{\text{obj}}$ is used to derive a geometry token, which are then jointly used to condition the diffusion model. Both tokens modulate the U-Net denoiser through cross-attention, enabling the diffusion model to generate an inpainted latent feature $\tilde{z}$ that is consistent with the transformation. The latent prediction is decoded using the pretrained image decoder D_{img}, yielding the final augmented image

$$x_{\text{aug}} = D_{\text{img}}(\tilde{z_0}). \tag{11}$$

It is paired with the transformed LiDAR points $\tilde{\mathcal{P}}_{\text{obj}}$ to form a geometry-aligned multi-modal augmentation sample, overcoming the misalignment and visual artifacts inherent in traditional object sampling and object-paste.

IV. Experiments

A. Experimental Settings

Datasets. The experiments are evaluated on the nuScenes [18], which provides 700 training scenes and 150 validation scenes. We restricted augmentation to the front cam attribute, since the diffusion models used in our work do not guarantee multi view consistency. Also, we focus on two object categories, car and pedestrian, which representative classes in driving scenes.

Multimodal Fusion Model. We used BEVFusion [4] as the 3D object detection model and evaluated augmentation performance using per-class mAP and NDS. To fairly measure augmentation performance, all detectors are trained from scratch, and each model is trained for a single object class. Image patches are resized to 512 × 512, LiDAR uses the standard multi-sweep setting, and each model is trained for about 48 hours on two NVIDIA A6000 GPUs.

Diffusion Model. The 3D-aware diffusion inpainting model is trained using cropped image patches of resolution 512 × 512. We fine-tune a Stable Diffusion v2.1 and use a latent resolution of 64 × 64. Generation takes 2.5 seconds per image, and training requires 24 hours on four NVIDIA A6000 GPUs with batch size 4.

Augmentation Strategy. To ensure a fair comparison across augmentation methods, we control the sampling procedure such that object sampling, object-paste, and our diffusion-based synthesis operate on the same object–image pairs. For each epoch, we record the indices of the target front camera image and the corresponding source object selected for augmentation. These index pairs are reused to generate the object-paste images and the diffusion-based inpainted results, so that all methods are evaluated under identical geometric configurations and object placements. Each epoch contains approximately 21,000 augmented front camera images, which are paired with the corresponding transformed LiDAR data and used to train the 3D detection model.

B. Qualitative Result

Figure 3 compares the visual results of the original images, the object-paste baseline, and our diffusion-based augmentation for both the car and pedestrian classes. The object-paste baseline frequently introduces visual artifacts such as

[Baseline]: Multi-modal BEVFusion

Class	Object Sample		Object Paste		Ours	
	mAP	NDS	mAP	NDS	mAP	NDS
Car	74.0	73.5	80.7 (+6.7)	74.2 (+0.7)	**81.5** (+7.5)	**79.2** (+5.7)
Pedestrian	63.9	69.1	65.1 (+1.2)	**71.1** (+2.0)	**65.9** (+2.0)	70.8 (+1.7)

TABLE I: Evaluation of BEVFusion [4] on the nuScenes [18] dataset under multiple augmentation schemes. The proposed LiDAR-Camera aligned strategy achieves superior performance compared to non-aligned methods.

color discontinuities, blurred boundaries, and illumination mismatches, which stem from directly inserting cropped objects into the target scene. A typical failure case appears in the fourth column of Figure 3, where a day-time pedestrian is placed into a night-time background, resulting in clear visual inconsistency.

In contrast, our method preserves geometric alignment between LiDAR and image domains while maintaining consistent appearance with the surrounding scene. The diffusion-generated objects are adapted to the local background context, which effectively suppresses boundary artifacts and satisfies the imposed 3D geometric constraints. As a result, the proposed approach produces visually more natural and geometrically consistent augmentations than conventional strategies.

C. Quantitative Result

The impact of our augmentation framework is summarized in Table I. Compared with object sampling (LiDAR-only) and object-paste, our method achieves improvements of $+6.7$ mAP and $+0.7$ NDS for the car class. This indicates that correcting cross-modal misalignment between LiDAR and image modalities plays an important role in improving detection performance. A similar trend is observed for the pedestrian class, where object-paste consistently outperforms object sampling.

While object-paste provides a clear improvement over object sampling, its performance remains below that of our method, indicating that paste-based augmentation alone is insufficient. For the car class, our approach achieves additional gains of $+1.2$ mAP and $+5.0$ NDS over object-paste. This suggests that both geometric alignment and appearance consistency are jointly required for stable multi-modal fusion. For the pedestrian class, although mAP improves, NDS remains relatively low, which is likely attributable to the limited image fidelity of small objects generated by the diffusion model. We believe that improvements in diffusion resolution and generation quality will further enhance the effectiveness of our augmentation framework for multi-modal fusion training.

V. Conclusion

In this work, we proposed a novel cross-modal consistent augmentation framework based on diffusion-based approach. By integrating our approach with conventional LiDAR-specific object sampling strategies, the proposed method effectively preserves the geometric alignment between LiDAR and image modalities. Through experiments on nuScenes 3D object detection, our method consistently outperforms conventional object sampling and object-paste techniques, demonstrating that LiDAR–camera alignment plays a critical role in multi-modal training. However, the proposed framework requires approximately 2.5 seconds per image for generation, which incurs a considerably higher computational cost compared to previous real-time augmentation methods. Besides, our method is limited in its applicability to multi-view and multi-object scenarios. Nevertheless, given the rapid advancement of diffusion models, we expect that these limitations can be substantially alleviated in future work. We expect that further improvements along this direction will positively contribute to the training of multi-modal fusion models.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2024-00409492).

References

[1] A. Xiao, J. Huang, D. Guan, K. Cui, S. Lu, and L. Shao, "Polarmix: A general data augmentation technique for lidar point clouds," 2022. [Online]. Available: https://arxiv.org/abs/2208.00223

[2] L. Kong, J. Ren, L. Pan, and Z. Liu, "Lasermix for semi-supervised lidar semantic segmentation," 2023. [Online]. Available: https://arxiv.org/abs/2207.00026

[3] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," 2020. [Online]. Available: https://arxiv.org/abs/2004.10934

[4] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," 2024. [Online]. Available: https://arxiv.org/abs/2205.13542

[5] Z. Su, H. Lu, S. Jiao, J. Xiao, Y. Wang, and X. Chen, "Efficient multimodal 3d object detector via instance-level contrastive distillation," 2025. [Online]. Available: https://arxiv.org/abs/2503.12914

[6] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

[7] C. Wang, C. Ma, M. Zhu, and X. Yang, "Pointaugmenting: Cross-modal augmentation for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 794–11 803.

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2022. [Online]. Available: https://arxiv.org/abs/2112.10752

[9] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," 2020. [Online]. Available: https://arxiv.org/abs/2006.11239

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014. [Online]. Available: https://arxiv.org/abs/1406.2661

[11] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2022. [Online]. Available: https://arxiv.org/abs/1312.6114

[12] R. Gao, K. Chen, E. Xie, L. Hong, Z. Li, D.-Y. Yeung, and Q. Xu, "Magicdrive: Street view generation with diverse 3d geometry control," 2024. [Online]. Available: https://arxiv.org/abs/2310.02601

[13] A. Buburuzan, A. Sharma, J. Redford, P. K. Dokania, and R. Mueller, "Mobi: Multimodal object inpainting using diffusion models," 2025. [Online]. Available: https://arxiv.org/abs/2501.03173

[14] B. Yang, S. Gu, B. Zhang, T. Zhang, X. Chen, X. Sun, D. Chen, and F. Wen, "Paint by example: Exemplar-based image editing with diffusion models," 2022. [Online]. Available: https://arxiv.org/abs/2211.13227

[15] Z. Wu, Y. Rubanova, R. Kabra, D. A. Hudson, I. Gilitschenski, Y. Aytar, S. van Steenkiste, K. Allen, and T. Kipf, "Neural Assets: 3d-aware multi-object scene synthesis with image diffusion models," in *Advances in Neural Information Processing Systems*, 2024.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015. [Online]. Available: https://arxiv.org/abs/1505.04597

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[18] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," 2020. [Online]. Available: https://arxiv.org/abs/1903.11027