♊GeminiFusion: Efficient Pixel-wise Multimodal Fusion for Vision Transformer

Ding Jia [*1] **Jianyuan Guo** [*2] **Kai Han** [3] **Han Wu** [2] **Chao Zhang** [1] **Chang Xu**[✉ 2] **Xinghao Chen**[✉ 3]

jiading@stu.pku.edu.cn; {jianyuan.guo,han.wu}@sydney.edu.au; kai.han@huawei.com; c.zhang@pku.edu.cn

Abstract

Cross-modal transformers have demonstrated superiority in various vision tasks by effectively integrating different modalities. This paper first critiques prior token exchange methods which replace less informative tokens with inter-modal features, and demonstrate exchange based methods underperform cross-attention mechanisms, while the computational demand of the latter inevitably restricts its use with longer sequences. To surmount the computational challenges, we propose *GeminiFusion*, a pixel-wise fusion approach that capitalizes on aligned cross-modal representations. *GeminiFusion* elegantly combines intra-modal and inter-modal attentions, dynamically integrating complementary information across modalities. We employ a layer-adaptive noise to adaptively control their interplay on a per-layer basis, thereby achieving a harmonized fusion process. Notably, *GeminiFusion* maintains linear complexity with respect to the number of input tokens, ensuring this multimodal framework operates with efficiency comparable to unimodal networks. Comprehensive evaluations across multimodal image-to-image translation, 3D object detection and arbitrary-modal semantic segmentation tasks, including RGB, depth, LiDAR, event data, etc. demonstrate the superior performance of our *GeminiFusion* against leading-edge techniques. The PyTorch code is available here.

1. Introduction

In light of the increasing availability of low-cost sensors, multimodal fusion which leverages data from various sources has emerged as a pivotal catalyst for advancing artificial intelligence-driven perception in vision (Smith & Gasser, 2005; Baltrušaitis et al., 2018; Guo et al., 2022a). This approach has demonstrated remarkable potential, surpassing the unimodal paradigm across various downstream tasks, including autonomous driving (Ha et al., 2017; Li et al., 2022), semantic segmentation (Ye et al., 2019; Cao et al., 2021), video captioning (Sun et al., 2019a; Lu et al., 2019) and visual question answering (Antol et al., 2015; Ben-Younes et al., 2017).

In the current literature, dominant paradigms for the multimodal fusion can be categorized into two ad-hoc schemes, *i.e.*, interaction-based fusion (Shvetsova et al., 2022; Nagrani et al., 2021; Zhang et al., 2023a) and exchange-based fusion (Wang et al., 2020c; 2022b; Zhu et al., 2023). In early interaction-based methods, a common practice involved directly concatenating tokens from different modalities (Su et al., 2019). This straightforward fusion approach neglects inter-modal interactions and sometimes leads to a poorer performance than single-modal counterparts (Wang et al., 2020b; 2022b). While cross-attention mechanisms are introduced as a solution, the quadratic complexity of the full attention with an increasing number of input tokens challenges the feasibility of cross-modal models. To tackle this issue, a simple strategy is to confine cross-modal interaction to later layers, often referred to as late-fusion (Nagrani et al., 2021). However, this method restricts the ability of the network's shallow layers to access valuable features from another modality, diminishing the original goal of facilitating mutual assistance between modalities and hindering overall model performance.

Exchange-based fusion provides a parameter-free solution (Wang et al., 2022b; 2020c) to the computational overhead by leveraging the inherent alignment of different modalities in vision tasks. For instance, world-space data like LiDAR and point clouds can be projected to pixels on the paired image plane. This method entails dynamically predicting the significance of each input token and subsequently replacing less crucial tokens from one modality with those from another.

Our investigation into the prune-then-substitute technique, as outlined in the TokenFusion (Wang et al., 2022b), reveals

[*] Equal contribution [1] Peking University. [2] The University of Sydney. [3] Huawei Noah's Ark Lab.. Correspondence to: Chang Xu <c.xu@sydney.edu.au>, Xinghao Chen <xinghao.chen@huawei.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

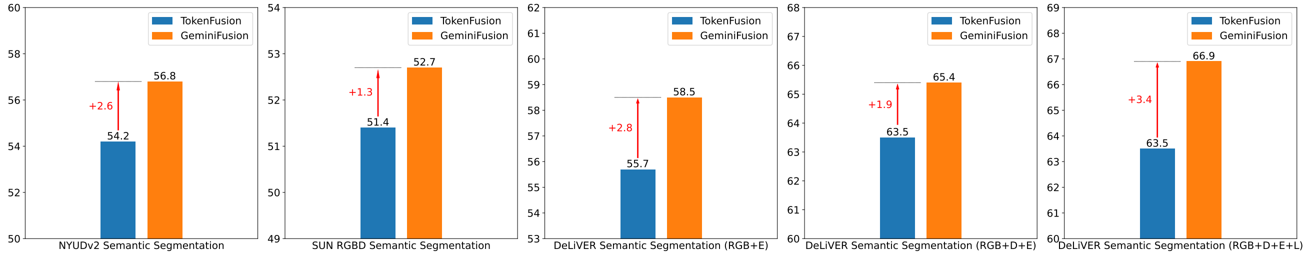


Figure 1: Improvements of our ♊GeminiFusion across five multimodal semantic segmentation tasks. GeminiFusion achieves +2.6%, +1.3%, +2.8%, +1.9%, and +3.4% performance gains. All training epoch numbers are aligned. D: Depth, E: Event, L: LiDAR.

that its effectiveness is not as consistent as expected. We observe that the network's shallow layers deem all tokens insignificant and indiscriminately substitute them with representations from an alternate modality. This behavior is in stark contrast to that of the deeper layers, which align more closely with our initial expectations by selectively swapping out representations of less pivotal tokens. Moreover, our results suggest that a strategy of unconditionally exchanging all tokens almost invariably yields the best outcomes, as evidenced by the data presented in Figure 3. Upon further analysis, we believe that this phenomenon can be attributed to the intrinsic unique information carried by each token; any direct substitution results in an irrevocable loss of information. We also note instances of simultaneous information exchange at identical positions across modalities, underscoring the necessity for features from different modalities to be mutually retained and integrated.

We observe that the performance of the exchange-based fusion consistently underperforms the cross-attention based fusion, while the additional overhead introduced by the full attention poses a significant challenge. To overcome this challenge and maintain the core information captured by the original unimodal learning, we introduce a pixel-wise multimodal fusion approach called GeminiFusion. Specifically, given two modalities, only the two matched tokens from corresponding modalities will participate in the fusion process. This fusion scheme has a minimal impact on the original unimodal representations, on account of the preservation of skip connections from the original inputs and the retention of self-consistent part during the fusion process. Meanwhile, the cross-modality part can significantly capture valuable multimodal information. The computational cost is minor since the pixel-wise attention is more compact compared to the full attention. Moreover, GeminiFusion demonstrates its superiority by allowing multimodal architectures to leverage parameters from unimodal pre-training, such as on the ImageNet dataset.

To verify the advantage of the proposed method, we consider extensive tasks including multimodal image-to-image translation, 3D object detection and arbitrary-modal semantic segmentation, *i.e.*, RGB, depth, events, and LiDAR, covering four multimodal benchmarks.

Our contributions in this paper include: (i) we empirically demonstrate that directly replacing features of one modality with those from another modality is sub-optimal. Simply exchanging all tokens every time achieves better results; (ii) we propose an efficient method named GeminiFusion for multimodal feature fusion, leveraging the inherent high alignment of different modal inputs in vision tasks while preserving the original unimodal features; (iii) extensive experiments on multimodal image-to-image translation, 3D object detection tasks and arbitrary-modal segmentation consistently affirm the effectiveness of our proposed GeminiFusion.

2. Related Work

The process of multimodal fusion involves leveraging diverse data sources to enhance associated details, surpassing the capabilities of their unimodal counterparts. Here, we delve into two prevailing fusion schemes and emphasizing their applicability in targeted multimodal vision tasks.

Interaction-based multimodal fusion. Early studies of interaction-based fusion (Snoek et al., 2005; Atrey et al., 2010; Bruni et al., 2014) categorizes the fusion strategy into three broad types: early (input-level), mid (feature-level) and late (decision-level) fusion. Early fusion methods (Zhao et al., 2020; Zhang et al., 2021a) directly fuse the inputs from different modalities through a single-stream network, performed by averaging (Hazirbas et al., 2017) or concatenating (Zhang & Funkhouser, 2018) along the input channels. However, the supervision signal is distant from the blended input, resulting in suboptimal results. Additionally, maintaining supervision for individual modalities is not feasible within this framework. Mid fusion (Lin et al., 2017; Chen et al., 2019; Fu et al., 2020; Ramachandram & Taylor, 2017; De Vries et al., 2017) harnesses individual CNN or transformer encoders for each modality to capture intricacies in their respective features (Xu et al., 2023; Guo et al., 2022b). For example, MBT(Nagrani et al., 2021)

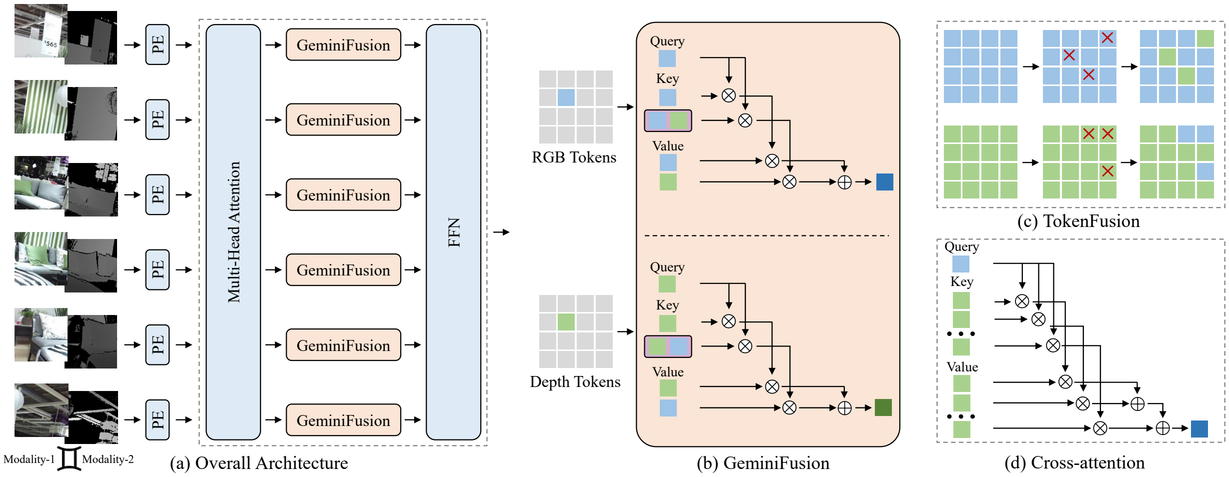


Figure 2: (a) Overall architecture of GeminiFusion: our proposed GeminiFusion model is designed to be plug and play, allowing it to be seamlessly integrated into various vision backbones. (b) GeminiFusion module: performing pixel-wise fusion to enrich multimodal feature by utilizing aligned features from two modalities. (c) TokenFusion: swapping certain pixels between two features, but result in information loss. (d) Cross-attention: requires a significant amount of memory resources with quadratic complexity of input token.

subsequently amalgamated the encoded features through a dedicated fusion layer. RDFNet (Park et al., 2017) and CMX (Zhang et al., 2023a) employ multilayer fusion, aggregating features iteratively with additional convolutional blocks. EPIC-Fusion (Kazakos et al., 2019) combines intermediate activations via summation in the joint training of multiple modality-specific networks. TransFuser (Prakash et al., 2021) utilizes several transformer modules for the fusion of intermediate features between different modalities. Late fusion (Owens & Efros, 2018) aggregates the final decision through an ensemble of multiple outputs (Pandeya & Lee, 2021; Glodek et al., 2011), usually implemented using parallel networks.

Exchange-based multimodal fusion. CEN (Wang et al., 2020c) introduces the parameter-free Channel Exchanging Network, which dynamically exchanges channels between sub-networks of different modalities. MLF-VO(Jiang et al., 2022) extends this method to fuse color and inferred depth maps, incorporating a polarization regularizer to prevent the model from reaching a singular solution. MuSE (Zhu et al., 2023) generalizes exchange-based methods from vision-vision fusion to text-vision fusion. TokenFusion (Wang et al., 2022b), on the other hand, performs the exchange in the token dimension. It dynamically detects uninformative tokens and substitutes these tokens with features from other modalities. In this paper, we contend that the prune-then-substitute approach employed by TokenFusion consistently falls short in performance compared to the cross-attention-based interaction method. There is also a risk that all tokens undergo unnecessary exchange, resulting in irreversible information loss.

Attention for multimodal fusion. Attention mechanisms, including self-/cross-attention (Vaswani et al., 2017), CBAM (Woo et al., 2018), SENet (Hu et al., 2018), and ECA (Wang et al., 2020a) have demonstrated their success in various tasks. Several multimodal frameworks (Li et al., 2022; Hori et al., 2017; Wei et al., 2020) incorporate attention modules to fuse features from different modalities. For instance, ACNet (Hu et al., 2019) processes RGB and depth with two branches and employs the proposed Attention Complementary Module (ACM) to enable the fusion branch, exploiting more high-quality features from different channels. Different from the ACNet, we concentrate more on the aligned spatial location to explore an efficient fusion method. VST (Liu et al., 2021a) utilizes cross-attention to fuse features from two modalities by computing the self-attention between the queries from one modality and the keys and values from the other modality. TransFuser (Prakash et al., 2021) and TriTransNet (Liu et al., 2021c) concatenate two modal features and use self-attention to mix information. Additionally, works like (Zhao et al., 2021; Wang et al., 2022a) employ the SE module to blend information. In contrast to previous quadratic complexity cross-attention, our pixel-wise attention has linear complexity with respect to the number of input tokens. This feature enables our fusion method to maintain a nearly as compact multimodal architecture as a unimodal network.

Multimodal semantic segmentation. Many segmentation methods excel in standard RGB-based benchmarks, providing per-pixel category predictions in a given scene. However, they often face challenges in real-world scenarios with rich 3D geometric information. To overcome this limitation, researchers have sought to enhance scene understanding by incorporating multimodal sensing, including depth (Silberman et al., 2012; Gupta et al., 2014), thermal (Ha et al., 2017; Sun et al., 2019b), polarimetric optical cues (Kalra et al., 2020), event-driven priors (Zhang et al., 2021b), and LiDAR (Zhuang et al., 2021a; Caesar et al., 2020). Previous works have primarily focused on the RGB-depth setting, which may not generalize well across different sensing data (Zhang et al., 2023a). In this study, we explore a unified approach capable of generalizing effectively to diverse multimodal combinations for semantic segmentation.

3. Our Method

We first revist the recently proposed TokenFusion (Wang et al., 2022b) method in Section 3.1. Subsequently, Section 3.2 details the commonly utilized cross-attention mechanism. Our pixel-wise GeminiFusion module is introduced in Section 3.3, and the comprehensive architecture is presented in Section 3.4.

3.1. Fusion via exchange

Based on the motivation that there are always uninformative tokens or channels in single-modal transformers, exchange based methods such as TokenFusion (Wang et al., 2022b) and CEN (Wang et al., 2020c) are designed to dynamically detect and substitute these useless tokens or channels with features from other modalities. Specifically, at the core of its functionality, TokenFusion (Wang et al., 2022b) prunes tokens in each modality and replaces them with corresponding tokens from other modalities that have been projected and aggregated to match. This exchange is guided by a score predictor integrated within each block of the network, which computes masks that share the dimensions of the multimodal inputs. These masks, through a comparison against a predefined threshold, facilitate the selection of tokens to be substituted. Specifically, if there are only two modalities as input, *i.e.*, $\mathbf{X}^1$ and $\mathbf{X}^2$, the token exchange process can be formulated as:

$$\begin{aligned}\mathbf{X}^1_{[\mathrm{i}]} &= \mathbf{X}^1_{[\mathrm{i}]} \odot \mathbb{I}_{s(\mathbf{X}^1_{[\mathrm{i}]})\geq\theta} + \mathbf{X}^2_{[\mathrm{i}]} \odot \mathbb{I}_{s(\mathbf{X}^1_{[\mathrm{i}]})<\theta}, \\ \mathbf{X}^2_{[\mathrm{i}]} &= \mathbf{X}^2_{[\mathrm{i}]} \odot \mathbb{I}_{s(\mathbf{X}^2_{[\mathrm{i}]})\geq\theta} + \mathbf{X}^1_{[\mathrm{i}]} \odot \mathbb{I}_{s(\mathbf{X}^2_{[\mathrm{i}]})<\theta}.\end{aligned} \tag{1}$$

where $\mathbf{X}^1_{[\mathrm{i}]}$ indicates the i-th token of input $\mathbf{X}^1$, $\mathbb{I}$ is an indicator asserting the subscript condition, therefore it outputs a mask tensor $\in \{0,1\}^N$, the parameter θ is a small threshold set to 0.02, and the operator $\odot$ resents the element-wise multiplication.

The supervision of the mask generation process is enforced through an L-1 norm constraint. However, this approach introduces an element of stochasticity. The model does not inherently prioritize the informational importance of tokens when generating the masks. We contend that the connection between the masks and the tokens' intrinsic information content is not well-regulated, which may lead to randomness in the exchange process. As demonstrated in Figure 3c and Figure 3d, altering the threshold does not prevent the tokens in the initial layers from being entirely exchanged. This suggests that TokenFusion does not operate as initially hoped, where tokens with negligible information are replaced by those from other modalities. Furthermore, as illustrated in Figure 3a and Figure 3b, setting the threshold to 1, thereby allowing all tokens always to be exchanged, yields better results. This indicates that the exchange-based method of TokenFusion is not only unstable but also prone to the loss of critical information. Hence, it may be less effective than a strategy involving the complete exchange of information.

Table 1: Comparison with TokenFusion on the NYUDv2, SUN RGB-D and DeLiVER datasets for multimodal semantic segmentation task. Evaluation metrics include pixel accuracy (%), mean accuracy (%), and mean IoU (%). Only mIoU is reported on the DeLiVER dataset following CMNeXt (Zhang et al., 2023b). † marks the methods are reproduced by ourselves. All training epochs are aligned. D: Depth, E: Event, L: LiDAR.

Method	Backbone	Inputs	Pixel Acc.	mAcc.	mIoU
Results on the NYUDv2 dataset					
TokenFusion	MiT-B3	RGB+D	79.0	66.9	54.2
GeminiFusion	MiT-B3	RGB+D	$\mathbf{79.9}^{+0.9}$	$\mathbf{69.9}^{+3.0}$	$\mathbf{56.8}^{+2.6}$
TokenFusion†	MiT-B5	RGB+D	79.1	67.5	55.1
GeminiFusion	MiT-B5	RGB+D	$\mathbf{80.3}^{+1.2}$	$\mathbf{70.4}^{+2.9}$	$\mathbf{57.7}^{+2.6}$
Results on the SUN RGB-D dataset					
TokenFusion†	MiT-B3	RGB+D	82.8	63.6	51.4
GeminiFusion	MiT-B3	RGB+D	$\mathbf{83.3}^{+0.5}$	$\mathbf{64.6}^{+1.0}$	$\mathbf{52.7}^{+1.3}$
TokenFusion†	MiT-B5	RGB+D	83.1	63.9	51.8
GeminiFusion	MiT-B5	RGB+D	$\mathbf{83.8}^{+0.7}$	$\mathbf{65.3}^{+1.4}$	$\mathbf{53.3}^{+1.5}$
Results on the DeLiVER dataset					
TokenFusion†	MiT-B2	RGB+D	-	-	63.7
GeminiFusion	MiT-B2	RGB+D	-	-	$\mathbf{66.4}^{+2.7}$
TokenFusion†	MiT-B2	RGB+E	-	-	55.7
GeminiFusion	MiT-B2	RGB+E	-	-	$\mathbf{58.5}^{+2.8}$
TokenFusion†	MiT-B2	RGB+L	-	-	55.5
GeminiFusion	MiT-B2	RGB+L	-	-	$\mathbf{58.6}^{+3.1}$
TokenFusion†	MiT-B2	RGB+D+E+L	-	-	63.5
GeminiFusion	MiT-B2	RGB+D+E+L	-	-	$\mathbf{66.9}^{+3.4}$

3.2. Fusion via cross-attention

We commence with an exploration of a prevalent cross-attention-based fusion architecture (Li et al., 2022; Carion et al., 2020), which is typified by the utilization of a canonical attention scheme to process inputs derived from multiple modalities. As illustrated in Figure 2d, consider the scenario where we have procured a set of N patches from two modalities, denoted as $\mathbf{X}^1, \mathbf{X}^2 \in \mathbb{R}^{N\times d}$, the corresponding output $\mathbf{Y}^1, \mathbf{Y}^2 \in \mathbb{R}^{N\times d}$ augmented by multimodal information can be generated by:

$$\begin{aligned}\mathbf{Y}^1 &= \mathrm{Attention}(\mathbf{X}^1\mathrm{W}^\mathrm{Q}, \mathbf{X}^2\mathrm{W}^\mathrm{K}, \mathbf{X}^2\mathrm{W}^\mathrm{V}) + \mathbf{X}^1 \\ \mathbf{Y}^2 &= \mathrm{Attention}(\mathbf{X}^2\mathrm{W}^\mathrm{Q}, \mathbf{X}^1\mathrm{W}^\mathrm{K}, \mathbf{X}^1\mathrm{W}^\mathrm{V}) + \mathbf{X}^2 \\ &\mathrm{Attention}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{Softmax}(\mathrm{QK}^\mathrm{T}/\sqrt{\mathrm{d}})\mathrm{V}\end{aligned} \tag{2}$$

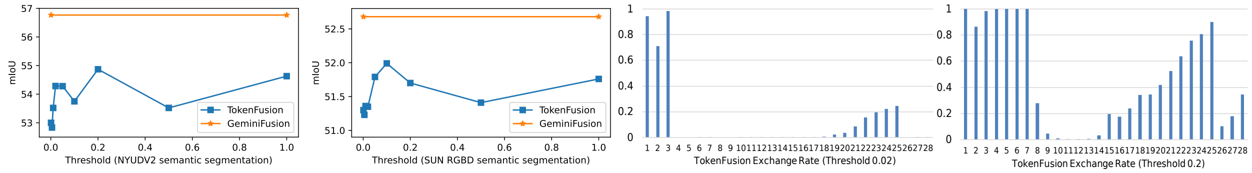


Figure 3: Impact of the threshold on the exchange-based TokenFusion. Exchanging all tokens almost invariably yields the best outcomes.

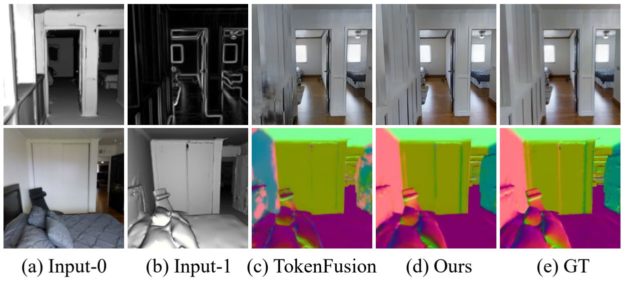


Figure 4: Image-to-image translation results on the validation split of Taskonomy. Best view in color and zoom in.

Table 2: Comparison on the Taskonomy dataset for the multimodal image-to-image translation task. Evaluation metrics are FID/KID (×10^{-2}) for the RGB predictions and MAE (×10^{-1})/MSE (×10^{-1}) for other predictions. Lower values indicate better performance for all the metrics. All training epoch numbers are aligned.

Method	Shade+Texture → RGB	Depth+Normal → RGB	RGB+Shade → Normal	RGB+Edge → Depth
TokenFusion	47.31/0.94	103.87/4.24	0.67/1.75	0.22/0.55
GeminiFusion	**41.32**$^{-5.99}$/**0.81**$^{-0.13}$	**96.98**$^{-6.89}$/**3.71**$^{-0.53}$	**0.65**$^{-0.02}$/**1.69**$^{-0.06}$	**0.20**$^{-0.02}$/**0.49**$^{-0.06}$

Table 3: Comparison with MVX-Net on the 3D object detection task against vehicle targets. The dataset is the validation set of the KITTI 3D object detection dataset. All training epoch numbers are aligned. The IoU threshold is 0.7.

Method	Param(M)	3D AP_{R11}			3D AP_{R40}		
		Easy	Medium	Hard	Easy	Medium	Hard
MVX-Net	33.8	87.49	77.04	74.54	88.41	**78.77**	74.27
MVX-Net + GeminiFusion	34.8	**88.49**	**77.36**	**74.61**	**89.43**	78.76	**74.46**

The computational complexity of above operation is $\mathcal{O}(N^2 \cdot c)$, where N is the number of tokens of both modalities. Given that N can be exceptionally large, the computational demand of the model is significantly increased. For instance, CMNeXt (Zhang et al., 2023b) partitions each modality input into $16,384$ patches. This partitioning leads to a computational requirement of over 17G FLOPs for just one instance of cross-attention, a figure that is prohibitive for practical model deployment.

3.3. ♊GeminiFusion: pixel-wise fusion module

To harness the benefits of modality fusion through cross-attention mechanism while circumventing the computational intensity that it entails, we introduce an innovative pixel-wise fusion module, termed the GeminiFusion module.

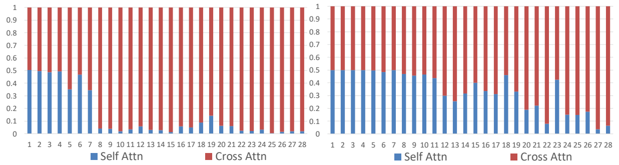


Figure 5: Comparison of attention scores obtained from self-attention (intra-modality) and cross-attention (inter-modality). Left: with noise. Right: without noise.

Drawing inspiration from TokenFusion (Wang et al., 2022b), we posit that not all patches contribute equally to the fusion process. Less salient patches could be efficiently substituted by their spatial counterparts from the alternate modality, implying that exhaustive interaction among all patches may not be obligatory. This insight leads us to the hypothesis that the crux of inter-modality information exchange lies in the patches sharing identical spatial coordinates, as these locations are where information exchange is most pertinent and significant. Leveraging this insight, the GeminiFusion module is engineered to prioritize interactions between spatially co-located patches from different modalities, thus refining the cross-attention mechanism:

$$\begin{aligned}\mathbf{Y}^1_{[\mathrm{i}]} &= \mathrm{Attention}(\mathbf{X}^1_{[\mathrm{i}]}\mathrm{W}^{\mathrm{Q}}, \mathbf{X}^2_{[\mathrm{i}]}\mathrm{W}^{\mathrm{K}}, \mathbf{X}^2_{[\mathrm{i}]}\mathrm{W}^{\mathrm{V}}) + \mathbf{X}^1_{[\mathrm{i}]},\\ \mathbf{Y}^2_{[\mathrm{i}]} &= \mathrm{Attention}(\mathbf{X}^2_{[\mathrm{i}]}\mathrm{W}^{\mathrm{Q}}, \mathbf{X}^1_{[\mathrm{i}]}\mathrm{W}^{\mathrm{K}}, \mathbf{X}^1_{[\mathrm{i}]}\mathrm{W}^{\mathrm{V}}) + \mathbf{X}^2_{[\mathrm{i}]}.\end{aligned} \tag{3}$$

where i is in the range of d. The targeted interaction strategy of GeminiFusion module not only focuses computational effort on the most critical information exchanges but also significantly slashes the computational load. This efficiency is quantified by a reduction in computational complexity to $\mathcal{O}(N \cdot c^2)$. Compared with the cross-attention, the FLOPs plummet from 17G to merely 0.14G. This staggering reduction of 99.2% in computational demand marks a transformative improvement, rendering the module exceedingly efficient for deployment in environments where computational resources are at a premium or where real-time performance is necessary.

However, two main challenges arise here: **(i) Incongruity outcomes from the attention score.** In the TokenFusion (Wang et al., 2022b) framework, the exchange of less informative patches with those from a different modality has been shown to enhance model performance. Conversely, within the attention module, a tendency arises where one modality disproportionately learns from patches of another modality that are more self-similar, as they yield higher attention scores. This proclivity is antithetical to our intended model behavior, which seeks to benefit from the integration of dissimilar and potentially more informative patch characteristics. **(ii) Softmax function limitation in per-pixel attention mechanism.** The current attention formulation operates on a per-pixel basis, resulting in an attention map of dimension 1×1. The application of the softmax function in this context is rendered ineffective as it invariably returns a value of one, nullifying the intended differentiation of the attention mechanism. This outcome undermines the capacity of the model to assign varying levels of attention across modalities.

To address the aforementioned issues, we propose two enhancements. Firstly, we introduce a lightweight *relation discriminator* to evaluate the disparity between modalities. Our findings indicate that a synergistic combination of a 1×1 convolution followed by a softmax function suffices. The associated experiments are detailed in Table 6. Specifically, patches from the two modalities are concatenated and fed into the relation discriminator, which subsequently assigns a relation score ranging from 0 to 1. This relation score is utilized to modulate the original key, effectively substituting the standard key in Eq. 2:

$$\begin{aligned}&\mathbf{Y}^1_{[\mathrm{i}]} = \mathrm{Attention}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) + \mathbf{X}^1_{[\mathrm{i}]} \\ &\mathrm{Q} = \mathbf{X}^1_{[\mathrm{i}]}\mathrm{W}^{\mathrm{Q}},\ \mathrm{K} = \mathbf{X}^1_{[\mathrm{i}]}\phi(\mathbf{X}^1_{[\mathrm{i}]}, \mathbf{X}^2_{[\mathrm{i}]})\mathrm{W}^{\mathrm{K}},\ \mathrm{V} = \mathbf{X}^2_{[\mathrm{i}]}\mathrm{W}^{\mathrm{V}}\end{aligned} \tag{4}$$

where $\phi(\cdot)$ indicates our relation discriminator module. The formula for $\mathrm{Y}^2_{[\mathrm{i}]}$ is obtained in the same way. To prevent the second issues associated with single-item focus without adding redundant information, we add the pixel-wise self-attention into the Eq. 4:

$$\begin{aligned}&\mathrm{K} = [\mathbf{X}^1_{[\mathrm{i}]}\mathrm{W}^{\mathrm{K}},\ \mathbf{X}^1_{[\mathrm{i}]}\phi(\mathbf{X}^1_{[\mathrm{i}]}, \mathbf{X}^2_{[\mathrm{i}]})\mathrm{W}^{\mathrm{K}}], \\ &\mathrm{V} = [\mathbf{X}^1_{[\mathrm{i}]}\mathrm{W}^{\mathrm{V}},\ \mathbf{X}^2_{[\mathrm{i}]}\mathrm{W}^{\mathrm{V}}].\end{aligned} \tag{5}$$

The formula for $\mathrm{Y}^2_{[\mathrm{i}]}$ is obtained in the same way. In the self-attention mechanism described by Equation 5, both the query and key are derived from identical modal inputs, leading to an inherent bias towards the self-referential component of the attention score. This can diminish the efficacy of learning cross-modal representations. To address this issue, we augment the self-attention with *layer-adaptive noise*. This approach involves the injection of a minimal amount of noise at the layer level, subtly enhancing the feature representation without burdening the model with extraneous information. To encapsulate this process for input tensors $\mathbf{X}^1, \mathbf{X}^2 \in \mathbb{R}^{N \times d}$ at Layer L, the resultant output tensors $\mathbf{Y}^1, \mathbf{Y}^2 \in \mathbb{R}^{N \times d}$ within our GeminiFusion module can be mathematically represented as follows:

$$\begin{aligned}\mathbf{Y}^1_{[\mathrm{i}]} &= \mathrm{Attention}(\mathrm{Q}^1, \mathrm{K}^1, \mathrm{V}^1) + \mathbf{X}^1_{[\mathrm{i}]}, \\ \mathrm{Q}^1 &= \mathbf{X}_{1[\mathrm{i}]}W^Q, \\ \mathrm{K}^1 &= [(\mathrm{Noise}^{\mathrm{K}}_{\mathrm{L}} + \mathbf{X}^1_{[\mathrm{i}]})\mathrm{W}^{\mathrm{K}},\ \mathbf{X}^1_{[\mathrm{i}]}\phi(\mathbf{X}^1_{[\mathrm{i}]}, \mathbf{X}^2_{[\mathrm{i}]})\mathrm{W}^{\mathrm{K}}], \\ \mathrm{V}^1 &= [(\mathrm{Noise}^{\mathrm{V}}_{\mathrm{L}} + \mathbf{X}^1_{[\mathrm{i}]})\mathrm{W}^{\mathrm{V}},\ \mathbf{X}^2_{[\mathrm{i}]}\mathrm{W}^{\mathrm{V}}] \\ \mathbf{Y}^2_{[\mathrm{i}]} &= \mathrm{Attention}(\mathrm{Q}^2, \mathrm{K}^2, \mathrm{V}^2) + \mathbf{X}^2_{[\mathrm{i}]}, \\ Q^2 &= \mathbf{X}^2_{[\mathrm{i}]}W^Q, \\ \mathrm{K}^2 &= [(\mathrm{Noise}^{\mathrm{K}}_{\mathrm{L}} + \mathbf{X}^2_{[\mathrm{i}]})\mathrm{W}^{\mathrm{K}},\ \mathbf{X}^2_{[\mathrm{i}]}\phi(\mathbf{X}^2_{[\mathrm{i}]}, \mathbf{X}^1_{[\mathrm{i}]})\mathrm{W}^{\mathrm{K}}], \\ \mathrm{V}^2 &= [(\mathrm{Noise}^{\mathrm{V}}_{\mathrm{L}} + \mathbf{X}^2_{[\mathrm{i}]})\mathrm{W}^{\mathrm{V}},\ \mathbf{X}^1_{[\mathrm{i}]}\mathrm{W}^{\mathrm{V}}].\end{aligned} \tag{6}$$

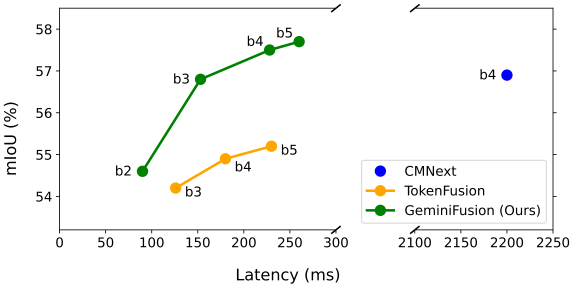


Figure 6: Performance *vs.*latency on the NYUDv2 dataset. GeminiFusion achieves the better trade-off compared with others. Latency is measured by averaging all validation samples of the NYUDv2 dataset. Multi-scale flip test strategy is used in CMNext as described in (Zhang et al., 2023b).

We have conducted an ablation study on noise selection, detailed in Table 7. Our findings indicate that the optimal noise implementation involves a learnable parameter added to the key, with this parameter being unique to each layer. This layer-specific noise facilitates a dynamic balance between self-attention and cross-modal attention and ensures the appropriate functioning of the softmax operation. Figure 5 illustrates the variation in attention scores across increasing layer depths.

3.4. Overall architecture

Our GeminiFusion model adopts an encoder-decoder architecture, with the encoder featuring a four-stage structure akin to the widely recognized SegFormer (Xie et al., 2021) for the extraction of hierarchical features. For conciseness, Figure 2 illustrates only the initial stage out of the four.

The primary focus lies in multimodal fusion based on visual data, encompassing modalities such as RGB, depth, event, and LiDAR. These modalities are inherently homogeneous, as they represent different visual perspectives of the same subject, and can be readily converted into image-like formats (Zhuang et al., 2021b; Zhang et al., 2023b). Within our framework, all modalities utilize shared parameters with the exception of the Layer Normalization (LN) layers, facilitating a uniform processing approach. More specifically, the

Table 4: Comparison of multimodal semantic segmentation results on NYUDv2 and SUN RGBD datasets with Swin (Liu et al., 2021b) and MiT-B3/B5 (Xie et al., 2021) as encoder models. All training epochs are aligned. Swin-Tiny-1k and Swin-Large-22k are pre-trained on the ImageNet-1K and ImageNet-22k, respectively.

Method	Encoder	Param(M)	NYUDv2 mIoU	SUNRGBD mIoU
GeminiFusion	MiT-B3	75.8	56.8	52.7
	MiT-B5	137.2	57.7	53.3
	Swin-Tiny-1k	52.0	52.2	50.2
	Swin-Large-22k	369.2	**60.2**	**54.6**

Table 5: Comparison results with state-of-the-art methods on the NYUDv2, SUN RGB-D and DeLiVER datasets for the multimodal semantic segmentation task. Additional strategies indicate that the method uses strategies other than ImageNet classification pre-training. For the DeLiVER dataset, we follow CMNeXt to use MiT-B2 as backbone for fair comparison. Therefore "MiT-B5 (MiT-B2)" indicates that we use MiT-B5 for NUYDv2 and SUN RGB-D, while MiT-B2 for DeLiVER. * indicates that we use the SUN RGBD trained model as pre-training on NYUDv2 dataset. † indicates that the results are reproduced by ourselves.

Method	Backbone	Additional Strategies	NYUDv2 mIoU	SUN RGBD mIoU	DeLiVER mIoU
PSD	ResNet50	✗	51.0	50.6	-
FSFNet	ResNet-101	✗	52.0	50.6	-
TokenFusion†	MiT-B5 (MiT-B2)	✗	55.1	51.8	63.5
SMMCL	SegNeXt-B	✗	55.8	-	-
MultiMAE	ViT-Base	✓	56.0	-	-
OMNIVORE	Swin-Large	✓	56.8	-	-
CMNeXt	MiT-B4 (MiT-B2)	✗	56.9	50.4	66.3
CMX	MiT-B5	✗	56.9	52.4	62.7
DFormer	DFormer-L	✓	57.2	52.5	-
PolyMaX	ConvNeXt-L	✓	58.1	-	-
SwinMTL	SwinV2-Base-MiM	✓	58.1	-	-
EMSANet	EMSANet-R34-NBt1D	✓	59.0	50.9	-
DPLNet	MiT-B5	✓	59.3	52.8	-
OmniVec	OmniVec-4	✓	60.8	-	-
GeminiFusion	MiT-B5 (MiT-B2)	✗	57.7	53.3	**66.9**
GeminiFusion	Swin-Large-22k	✗	60.2	**54.6**	-
GeminiFusion*	Swin-Large-22k	✓	**60.9**	-	-

RGB image $\boldsymbol{I}_{RGB} \in \mathcal{R}^{3\times H\times W}$, along with the other $M-1$ modalities $\boldsymbol{I}_{depth}, \cdots, \boldsymbol{I}_{LiDAR} \in \mathcal{R}^{3\times H\times W}$, undergoes sequential refinement through Multi-Head Self-Attention (MHSA) and Feed-Forward Network (FFN) blocks. These modalities are then adeptly integrated to harness intra-modal information via our proposed GeminiFusion module.

Upon completion of the four encoding stages, we obtain M sets of feature maps at different stages, denoted as $\boldsymbol{f}_l^m \in \{\boldsymbol{f}_1^m, \boldsymbol{f}_2^m, \boldsymbol{f}_3^m, \boldsymbol{f}_4^m\}$ for each modality $m \in [0, M-1]$. For the l-th encoding stage, the number of blocks per branch is specified by $b_l \in \{4, 8, 16, 32\}$, the stride by $s_l \in \{4, 8, 16, 32\}$, and the channel dimension by $C_l \in \{64, 128, 320, 512\}$. Within each stage, the M feature maps are fused into a singular feature map $\boldsymbol{f}$ through a process of weighted summation. Following the encoding process, the resultant four-stage features $\boldsymbol{f}_l \in \{\boldsymbol{f}_1, \boldsymbol{f}_2, \boldsymbol{f}_3, \boldsymbol{f}_4\}$ are channeled into the decoder. The decoder is responsible for synthesizing the segmentation predictions. We employ an MLP-based decoder, as outlined in SegFormer (Xie et al., 2021), to serve as our segmentation head.

By employing a single-branch design, we not only streamline network complexity but also enhance predictive generalization capabilities. Moreover, the shared parameter strategy aids in the detection of common patterns across different modalities, which is a key objective of multimodal fusion. It should be noted that while our method excels in processing homogeneous modalities where each data type represents a different perspective of the same input, it currently does not accommodate heterogeneous data combinations, such as images paired with audio or text. We also need to pre-define the method for aligning with the above data pairs. Addressing this limitation remains an avenue for future research.

4. Experiment

4.1. Datasets

For multimodal semantic segmentation experiments, we use the following datasets: **NYUDv2** (Silberman et al., 2012) dataset provides 795 training and 654 testing images, labeled into 40 categories. The resolution we use is 480x640, which is aligned with the setting in CMNeXt (Zhang et al., 2023b) and DFormer (Yin et al., 2023). **DeLiVER** (Zhang et al., 2023b) dataset contains 3983 training and 2005 testing images, which is more than four times the size of the NYUDv2 dataset. It has 25 classes. The resolution we use is 1024x1024, which is also aligned with CMNeXt. According to CMNext, only mIoU is reported. Thus, we also only report mIoU in experiments on the DeLiVER dataset. **SUN RGB-D** (Song et al., 2015) dataset contains 5285 training and 5050 testing images, which is about seven times the size of the NYUDv2 dataset and 1.7 times the size of the DeLiVER dataset. The input resolution is 480x480, which is aligned with DFormer. The class number of the SUN RGB-D dataset is 37.

For the image-to-image translation task, we follow the experiment settings used in CEN (Wang et al., 2020c) and TokenFusion (Wang et al., 2022b). **Taskonomy** (Zamir et al., 2018) dataset is a large-scale indoor scene dataset, which contains about 4 million indoor images. More than 10 modals are provided with each image, like depth, normal, shade, texture and edge. Each modal is of size 512x512. We use the same sampling strategy with CEN and TokenFusion, which takes 1000 training and 500 testing images. Our implementation details can be found in the appendix A.

For the 3D object detection task, we follow the experiment settings used in MVX-Net (Sindagi et al., 2019). **KITTI 3D object detection** (Geiger et al., 2012) dataset contains 7481 training samples and 7518 test samples. The test difficulty is categorized into three levels: easy, medium and hard, which is based on the size of the object, the degree of visibility (occlusion), and the degree of truncation. In this paper, like MVX-Net (Sindagi et al., 2019), the training set is further split into a training set and a validation set. After splitting,

Table 6: Ablation about the relation discriminator on the NYUDv2 dataset. All training epoch numbers are aligned. We use the MiT-B3 as the backbone.

Relation Discriminator	Pixel Acc.	mAcc.	mIoU
2layer-MLP	79.3	69.1	55.7
2layer-MLP + Sigmoid	79.5	69.7	55.9
2layer-MLP + Softmax	**79.9**	**69.9**	**56.8**
1x1CNN + Softmax	79.2	69.2	55.7
3x3CNN + 1x1CNN + Softmax	79.4	69.6	55.6
5x5CNN + 3x3CNN + 1x1CNN + Softmax	79.1	68.7	54.9
5x5CNN + 3x3CNN + 1x1CNN + 2layer-MLP + Softmax	79.2	68.9	55.3

Table 7: Ablation about the noise selection on the NYUDv2 dataset. All training epoch numbers are aligned. We use the MiT-B3 backbone.

Noise type	Pixel Acc.	mAcc.	mIoU
Random Gaussian Noise, Multiply	79.2	69.3	55.5
Random Gaussian Noise, Add	79.2	68.8	55.3
Learnable parameter, Multiply	79.6	69.2	56.2
Learnable parameter, Add	**79.9**	**69.9**	**56.8**

Table 8: Multimodal semantic segmentation results on NYUDv2 and SUN RGB-D datasets by adding our GeminiFusion only to last k layers. All models use the MiT-B3 backbone. All training epoch numbers are aligned. Latency is measured by averaging all validation samples in the NYUDv2 dataset.

Method	k	Param(M)	GFLOPs	Latency(ms)	NYUDv2 mIoU	SUN RGB-D mIoU
TokenFusion	28	45.9	108	126	54.2	51.4
GeminiFusion	28	75.8	174	153	**56.8**	**52.7**
GeminiFusion	22	75.1	165	144	**56.5**	**52.5**
GeminiFusion	16	69.3	152	129	**56.4**	**52.5**
GeminiFusion	10	62.5	138	**116**	**56.4**	**52.2**
GeminiFusion	4	55.7	124	**103**	**56.1**	**51.9**
GeminiFusion	1	48.8	119	**102**	**55.1**	**51.9**
GeminiFusion	0	45.9	108	**95**	53.3	51.2

Table 9: Ablation about different parts of GeminiFusion on the NYUDv2 dataset. PWC: point-wise cross-attention, NSA: noised self-attention, ARD: attention relation discriminator.

PWC	NSA	ARD	mIoU
✗	✗	✗	53.3
✓	✗	✗	55.4
✓	✓	✗	56.3
✓	✓	✓	**56.8**

the training set consists of 3712 samples and the validation set consists of 3769 samples.

4.2. Comparisons with TokenFusion

Table 1 summarizes the comparative analysis between GeminiFusion and TokenFusion on segmentation tasks. Overall, with consistent training and testing conditions, GeminiFusion outperforms TokenFusion across the board when it comes to the fusion of two to four modalities. Specifically, in scenarios where RGB is fused with Depth, GeminiFusion achieves an improvement of approximately 1%-2.6% over TokenFusion. When all four modalities are fused, GeminiFusion further extends its lead by a significant margin of 3.4% in mIoU, underscoring the efficacy of our attention-based fusion approach that retains essential information without loss.

Table 2 presents the corresponding results for the image-to-image translation task. Our GeminiFusion outstrips TokenFusion across all evaluated settings. For instance, in the Shade+Texture→RGB task, GeminiFusion attains FID/KID scores of 41.32/0.81, which is notably superior to TokenFusion with a relative decrease of 12.6% in the FID metric. Qualitative results, as illustrated in Figure 4, reveal that predictions using our GeminiFusion exhibit more natural patterns and are smoother and clearer in terms of colors and details. This demonstrates GeminiFusion's capability to preserve a more complete spectrum of the shade information.

4.3. Applying to Swin Transformer

The proposed GeminiFusion module is a plug-and-play module that can be inserted into existing multimodal architectures (predominantly into encoders) for enhancing the model's cross-modal learning capabilities. This modular approach allows GeminiFusion to take advantage of different architectures to improve the model's performance in multimodal tasks. In the previous experiments, we follow the TokenFusion codebase, which uses the SegFormer (Xie et al., 2021) as the encoder and a simple FFN as the decoder. However, in addition to SegFormer, models such as Swin Transformer (Liu et al., 2021b) can also be used as encoder models, which together with the decoder form a complete segmentation model. We further conducts several experiments on the Swin Transformer. Specifically, we inserts GeminiFusion into the SwinBlock. The official checkpoints of Swin Transformer pre-trained on the ImageNet classification task can also be loaded directly without degradation of accuracy, which demonstrates the advantages of our approach. The experimental results are shown in Table 4. It can be seen that GeminiFusion is also applicable in frameworks such as Swin Transformer, and in the case of using the Swin-Large-22k model, which was pre-trained on a larger ImageNet-22k dataset and with a larger number of parameters, as the baseline model, GeminiFusion also achieves optimal results among encoders, which reflects the plug-and-play nature of GeminiFusion in different frameworks, as well as its ability to successfully leverage the better representational capabilities provided by larger encoders.

4.4. Comparisons with state-of-the-art methods

In this paper, GeminiFusion is benchmarked against state-of-the-art multimodal segmentation methods on NYUDv2, SUN RGB-D, and DeLiVER datasets, and the results are detailed in Table 5. To ensure the fairness of the comparison, all methods that use pre-training methods and training strategies other than pre-training on the ImageNet classification tasks are labeled as "Additional Strategies", such as PolyMax (Yang et al., 2024) (pre-training is performed using ImageNet-22K and Taskonomy), DPLNet (Dong et al., 2023) (using pre-trained segmentation model), OmniVec (Srivastava & Sharma, 2024) (pre-trained based on self-supervision of large-scale masks), DFormer (Yin et al., 2023) (utilizes an RGB-D pre-trained backbone), EMSANet (Seichter et al., 2023) and OMNIVORE (Girdhar et al., 2022) (both of which utilize a strategy of multi dataset pre-training coupled with fine-tuning of individual datasets). In particular, we likewise attempts an additional pre-training strategy, using a GeminiFusion model (Swin Large-22k backbone) trained on the SUN RGBD dataset and fine-tuned on the NYUDv2 dataset.

As can be seen from the experimental results, GeminiFusion using the Swin-Large-22k backbone network achieves the highest level of performance on both NYUDv2 and SUN RGB-D datasets. Moreover, when fusing modalities such as RGB with Depth, Event and LiDAR data, GeminiFusion with the MiT-B2 backbone secures substantial gains over CMNeXt, attesting to the efficacy of our pixel-wise fusion methodology in handling highly aligned modalities. Additionally, we juxtapose the performance of the MiT-B4-based GeminiFusion with CMNeXt on the NYUDv2 dataset, as illustrated in Figure 6. Here, GeminiFusion not only attains marginally superior results but also boasts significantly reduced latency, even in the absence of multi-scale and flip testing augmentations typically employed by CMNeXt.

4.5. Effect of each component on GeminiFusion

We present an ablation study on the NYUDv2 dataset to assess the contribution of each component within our GeminiFusion framework. Table 9 shows our implementation of point-wise cross-attention yields a 2.1% increase in mIoU compared to the baseline, demonstrating that direct information exchange between modalities can lead to substantial gains. Additionally, the effectiveness of the noise-adaptive self-attention mechanism is evidenced by its ability to preserve intra-modal features, thereby preventing the loss of valuable information. The proposed relation discriminator can help refine the generation process of key features within the attention mechanism, ensuring more precise adjustments that improve overall performance.

4.6. Discussion on Inference Latency

Contrary to the TokenFusion approach as documented in (Wang et al., 2022b), our GeminiFusion method does not require integration at every layer within the network architecture. As evidenced by the experiments in Table 8, implementing GeminiFusion in only the final 10 layers still yields faster inference speeds while preserving accuracy, outperforming the benchmark method. Incorporating GeminiFusion even in just the last layer alone surpasses TokenFusion in terms of both inference latency and accuracy. However, it should be noted that the optimal results are achieved when GeminiFusion is applied across every layer.

Figure 6 graphically represents the trade-off between performance and latency. The comparison clearly demonstrates that our GeminiFusion significantly outperforms TokenFusion in terms of efficiency by a considerable margin.

4.7. 3D Object Detection task

We choose the MVX-Net (Sindagi et al., 2019) framework and the KITTI dataset for our 3D object detection experiments for vehicles. The experiments use images and depth maps as inputs for the detection of vehicle categories in the KITTI dataset, which is aligned with other works (Zhang et al., 2023c; Zheng et al., 2021). For the processing of the KITTI dataset, we choose the same dataset division and data processing methods as MVX-Net. GeminiFusion is inserted into the original fusion layer of MVX-Net, and the experimental results are shown in Table 3, which show that GeminiFusion achieves significant improvement in most of the performance indexes with almost no increase in the number of parameters, and a few performance indexes are almost the same as the benchmark model.

5. Conclusion

In this paper, we comprehensively examine exchange-based cross-modal transformers and point out their intrinsic deficiency in achieving comparable performance of cross-attention mechanisms. Furthermore, we propose a pixel-wise fusion approach named GeminiFusion, combining intra-modality and inter-modality attention for dynamic integration of complementary information across modalities. GeminiFusion achieves state-of-the-art performance across various multimodal semantic segmentation benchmark datasets, and also proved its effectiveness on image-to-image translation and 3D object detection tasks. It is worth noting that GeminiFusion operates with linear complexity with respect to the number of input tokens, achieving efficiency comparable with unimodal counterparts.

Acknowledgements

Ding Jia and Chao Zhang are supported by the National Nature Science Foundation of China under Grant 62071013 and 61671027, and National Key R&D Program of China under Grant 2018AAA0100300. Chang Xu is supported in part by the Australian Research Council under Projects DP240101848 and FT230100549.

Impact Statement

This paper contributes to the advancement of multimodal feature fusion in Machine Learning by comparing exchange-based fusion with cross-attention based fusion. Our findings consistently demonstrate that cross-attention based fusion outperforms exchange-based fusion by effectively preserving core information among features from different modalities. Additionally, we propose an efficient GenimiFusion approach to reduce the computational overhead associated with cross-attention. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2015.

Atrey, P. K., Hossain, M. A., El Saddik, A., and Kankanhalli, M. S. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 2010.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

Ben-Younes, H., Cadene, R., Cord, M., and Thome, N. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2017.

Bruni, E., Tran, N.-K., and Baroni, M. Multimodal distributional semantics. *Journal of artificial intelligence research*, 2014.

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and Beijbom, O. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

Cao, J., Leng, H., Lischinski, D., Cohen-Or, D., Tu, C., and Li, Y. Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. End-to-end object detection with transformers. In *European conference on computer vision*, 2020.

Chen, C., Rosa, S., Miao, Y., Lu, C. X., Wu, W., Markham, A., and Trigoni, N. Selective sensor fusion for neural visual-inertial odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

De Vries, H., Strub, F., Mary, J., Larochelle, H., Pietquin, O., and Courville, A. C. Modulating early visual processing by language. *Advances in Neural Information Processing Systems*, 2017.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dong, S., Feng, Y., Yang, Q., Huang, Y., Liu, D., and Fan, H. Efficient multimodal semantic segmentation via dual-prompt learning. *arXiv preprint arXiv:2312.00360*, 2023.

Fu, K., Fan, D.-P., Ji, G.-P., and Zhao, Q. Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

Geiger, A., Lenz, P., and Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Girdhar, R., Singh, M., Ravi, N., van der Maaten, L., Joulin, A., and Misra, I. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16102–16112, 2022.

Glodek, M., Tschechne, S., Layher, G., Schels, M., Brosch, T., Scherer, S., Kächele, M., Schmidt, M., Neumann, H., Palm, G., et al. Multiple classifier systems for the classification of audio-visual emotional states. In *Affective Computing and Intelligent Interaction: Fourth International Conference, ACII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part II*, 2011.

Guo, J., Han, K., Wu, H., Tang, Y., Chen, X., Wang, Y., and Xu, C. Cmt: Convolutional neural networks meet vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022a.

Guo, J., Tang, Y., Han, K., Chen, X., Wu, H., Xu, C., Xu, C., and Wang, Y. Hire-mlp: Vision mlp via hierarchical rearrangement. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2022b.

Gupta, S., Girshick, R., Arbeláez, P., and Malik, J. Learning rich features from rgb-d images for object detection and segmentation. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, 2014.

Ha, Q., Watanabe, K., Karasawa, T., Ushiku, Y., and Harada, T. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017.

Hazirbas, C., Ma, L., Domokos, C., and Cremers, D. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part I 13*, 2017.

Hori, C., Hori, T., Lee, T.-Y., Zhang, Z., Harsham, B., Hershey, J. R., Marks, T. K., and Sumi, K. Attention-based multimodal fusion for video description. In *Proceedings of the IEEE international conference on computer vision*, 2017.

Hu, J., Shen, L., and Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

Hu, X., Yang, K., Fei, L., and Wang, K. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In *2019 IEEE International conference on image processing (ICIP)*, pp. 1440–1444. IEEE, 2019.

Jiang, Z., Taira, H., Miyashita, N., and Okutomi, M. Self-supervised ego-motion estimation based on multi-layer fusion of rgb and inferred depth. In *2022 International Conference on Robotics and Automation (ICRA)*, 2022.

Kalra, A., Taamazyan, V., Rao, S. K., Venkataraman, K., Raskar, R., and Kadambi, A. Deep polarization cues for transparent object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Kazakos, E., Nagrani, A., Zisserman, A., and Damen, D. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., and Dai, J. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, 2022.

Lin, D., Chen, G., Cohen-Or, D., Heng, P.-A., and Huang, H. Cascaded feature network for semantic segmentation of rgb-d images. In *Proceedings of the IEEE international conference on computer vision*, 2017.

Liu, N., Zhang, N., Wan, K., Shao, L., and Han, J. Visual saliency transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021a.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021b.

Liu, Z., Wang, Y., Tu, Z., Xiao, Y., and Tang, B. Tritransnet: Rgb-d salient object detection with a triplet transformer embedding network. In *Proceedings of the 29th ACM international conference on multimedia*, 2021c.

Lu, J., Batra, D., Parikh, D., and Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 2019.

Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., and Sun, C. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 2021.

Owens, A. and Efros, A. A. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European conference on computer vision (ECCV)*, 2018.

Pandeya, Y. R. and Lee, J. Deep learning-based late fusion of multimodal information for emotion classification of music video. *Multimedia Tools and Applications*, 2021.

Park, S.-J., Hong, K.-S., and Lee, S. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, 2017.

Prakash, A., Chitta, K., and Geiger, A. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Ramachandram, D. and Taylor, G. W. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 2017.

Seichter, D., Stephan, B., Fischedick, S. B., Müller, S., Rabes, L., and Gross, H.-M. Panopticndt: Efficient and robust panoptic mapping. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7233–7240. IEEE, 2023.

Shvetsova, N., Chen, B., Rouditchenko, A., Thomas, S., Kingsbury, B., Feris, R. S., Harwath, D., Glass, J., and Kuehne, H. Everything at once-multi-modal fusion transformer for video retrieval. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2022.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, 2012.

Sindagi, V. A., Zhou, Y., and Tuzel, O. Mvx-net: Multimodal voxelnet for 3d object detection. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 7276–7282. IEEE, 2019.

Smith, L. and Gasser, M. The development of embodied cognition: Six lessons from babies. *Artificial life*, 2005.

Snoek, C. G., Worring, M., and Smeulders, A. W. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, 2005.

Song, S., Lichtenberg, S. P., and Xiao, J. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 567–576, 2015.

Srivastava, S. and Sharma, G. Omnivec: Learning robust representations with cross modal sharing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1236–1248, 2024.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., and Dai, J. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

Sun, C., Myers, A., Vondrick, C., Murphy, K., and Schmid, C. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019a.

Sun, Y., Zuo, W., and Liu, M. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, 2019b.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 2017.

Wang, F., Pan, J., Xu, S., and Tang, J. Learning discriminative cross-modality features for rgb-d saliency detection. *IEEE Transactions on Image Processing*, 2022a.

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., and Hu, Q. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020a.

Wang, W., Tran, D., and Feiszli, M. What makes training multimodal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020b.

Wang, Y., Huang, W., Sun, F., Xu, T., Rong, Y., and Huang, J. Deep multimodal fusion by channel exchanging. *Advances in neural information processing systems*, 2020c.

Wang, Y., Chen, X., Cao, L., Huang, W., Sun, F., and Wang, Y. Multimodal token fusion for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022b.

Wei, X., Zhang, T., Li, Y., Zhang, Y., and Wu, F. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 2018.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 2021.

Xu, Y., Li, C., Li, D., Sheng, X., Jiang, F., Tian, L., and Sirasao, A. Fdvit: Improve the hierarchical architecture of vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Yang, X., Yuan, L., Wilber, K., Sharma, A., Gu, X., Qiao, S., Debats, S., Wang, H., Adam, H., Sirotenko, M., et al. Polymax: General dense prediction with mask transformer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1050–1061, 2024.

Ye, L., Rochan, M., Liu, Z., and Wang, Y. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019.

Yin, B., Zhang, X., Li, Z., Liu, L., Cheng, M.-M., and Hou, Q. Dformer: Rethinking rgbd representation learning for semantic segmentation. *arXiv preprint arXiv:2309.09668*, 2023.

Zamir, A. R., Sax, A., Shen, W., Guibas, L. J., Malik, J., and Savarese, S. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3712–3722, 2018.

Zhang, J., Fan, D.-P., Dai, Y., Anwar, S., Saleh, F., Aliakbarian, S., and Barnes, N. Uncertainty inspired rgb-d saliency detection. *IEEE transactions on pattern analysis and machine intelligence*, 2021a.

Zhang, J., Yang, K., and Stiefelhagen, R. Exploring event-driven dynamic context for accident scene segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 2021b.

Zhang, J., Liu, H., Yang, K., Hu, X., Liu, R., and Stiefelhagen, R. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems*, 2023a.

Zhang, J., Liu, R., Shi, H., Yang, K., Reiß, S., Peng, K., Fu, H., Wang, K., and Stiefelhagen, R. Delivering arbitrary-modal semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023b.

Zhang, Y. and Funkhouser, T. Deep depth completion of a single rgb-d image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

Zhang, Y., Zhang, Q., Zhu, Z., Hou, J., and Yuan, Y. Glenet: Boosting 3d object detectors with generative label uncertainty estimation. *International Journal of Computer Vision*, 131(12): 3332–3352, 2023c.

Zhao, X., Zhang, L., Pang, Y., Lu, H., and Zhang, L. A single stream network for robust and real-time rgb-d salient object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, 2020.

Zhao, Y., Zhao, J., Li, J., and Chen, X. Rgb-d salient object detection with ubiquitous target awareness. *IEEE Transactions on Image Processing*, 2021.

Zheng, W., Tang, W., Jiang, L., and Fu, C.-W. Se-ssd: Self-ensembling single-stage object detector from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14494–14503, 2021.

Zhu, R., Han, C., Qian, Y., Sun, Q., Li, X., Gao, M., Cao, X., and Xian, Y. Exchanging-based multimodal fusion with transformer. *arXiv preprint arXiv:2309.02190*, 2023.

Zhuang, Z., Li, R., Jia, K., Wang, Q., Li, Y., and Tan, M. Perception-aware multi-sensor fusion for 3d lidar semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021a.

Zhuang, Z., Li, R., Jia, K., Wang, Q., Li, Y., and Tan, M. Perception-aware multi-sensor fusion for 3D LiDAR semantic segmentation. In *In Proceedings of the IEEE international conference on computer vision*, 2021b.

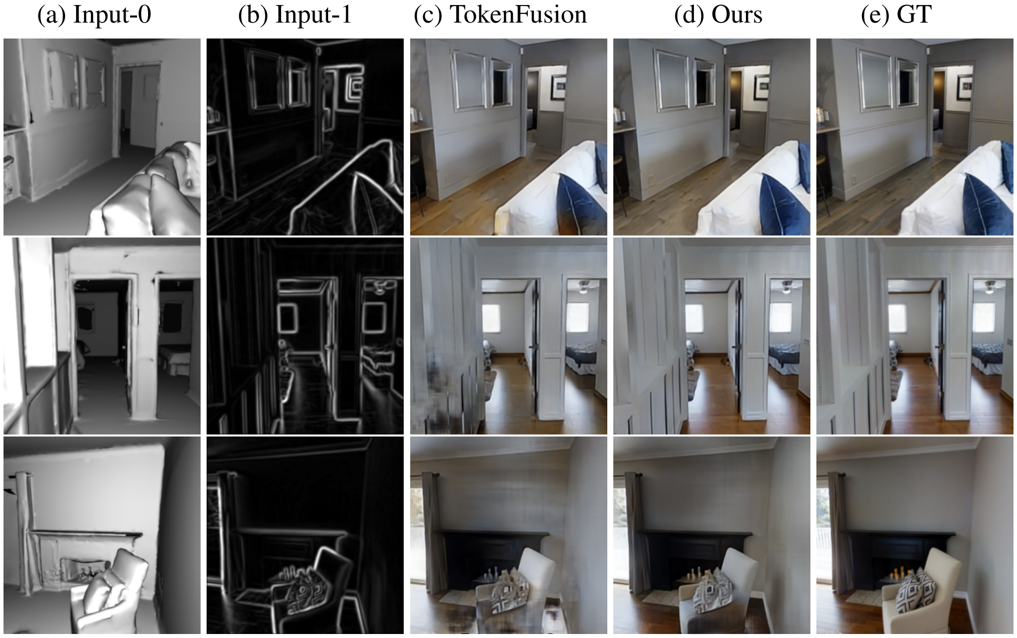
A. Implementation Details

- In the context of **multimodal semantic segmentation**, our training hyper-parameters are developed by following the methodologies from the TokenFusion (Wang et al., 2022b) and CMNeXt (Zhang et al., 2023b) codebases. For model training, we employ NVIDIA V100 GPUs in configurations of 3, 4, and 8 units for the NYUDv2, SUN RGB-D, and DeLiVER datasets, respectively, adhering to the same environmental settings as specified in the original papers. Our encoder design is an adaptation from SegFormer (Xie et al., 2021), which has been pre-trained solely on the ImageNet-1K (Deng et al., 2009) dataset for classification tasks. For experiments on the NYUDv2 and SUN RGB-D datasets, we utilize the setup from the TokenFusion, maintaining consistency in batch size, optimizer, learning rate, and learning rate scheduler. Within our proposed GeminiFusion model, we configure the number of attention heads to 8. To mitigate the risk of overfitting, we set the drop path rate to 0.4, while the drop rate remains at 0.0. Conversely, for the DeLiVER dataset, our foundation training hyper-parameters are the same with CMNeXt, which necessitates a smaller backbone. Consequently, we reduce the drop path rate to 0.2. All other parameters, including batch size, optimizer, weight decay, and learning rate scheduler, remain in line with CMNeXt's original configuration, except for the learning rate, which is modified to $2e^{-4}$.

- For the **image-to-image translation** task, we also follow the setting in TokenFusion and set the same hyper-parameters as the TokenFusion. We use one NVIDIA V100 card for all image-to-image translation experiments.

- For the **3D object detection** task, we also follow the setting in MVX-Net and set the same hyper-parameters as the MVX-Net. We use 4 NVIDIA V100 cards for all experiments.

B. More Visualization Results



(a) Input-0 (b) Input-1 (c) TokenFusion (d) Ours (e) GT

Figure 7: Shade+Texture→RGB. Image-to-image translation results on the validation split of Taskonomy (Zamir et al., 2018).

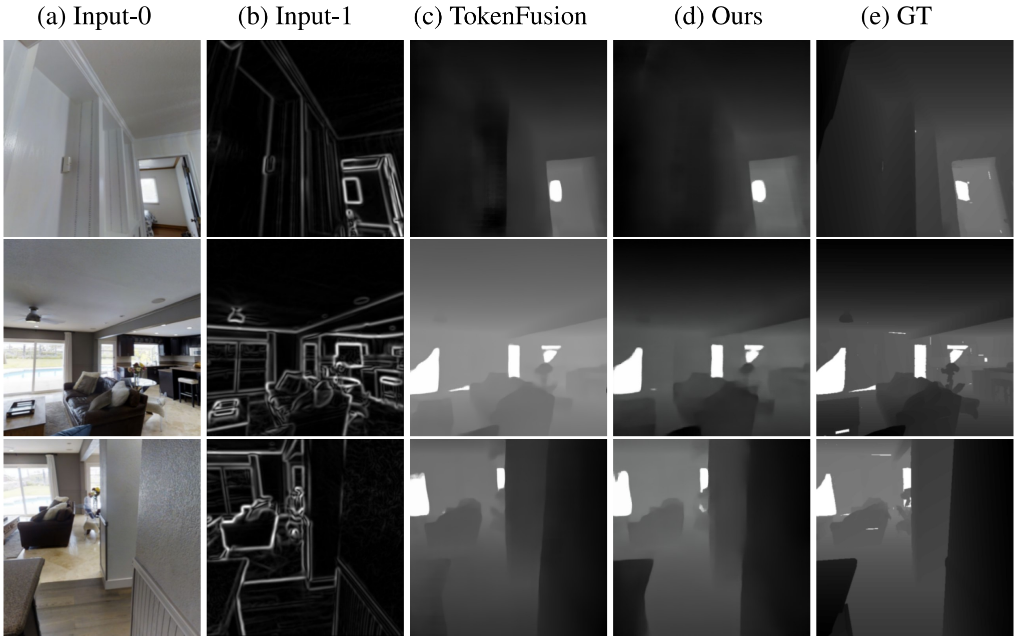


Figure 8: RGB+Edge→Depth. Image-to-image translation results on the validation split of Taskonomy.

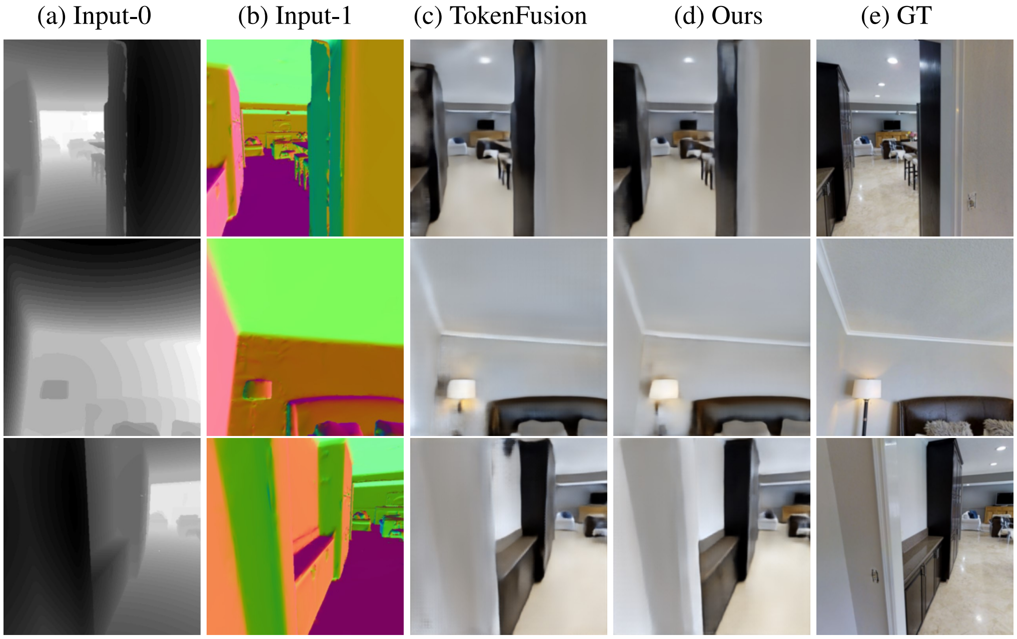


Figure 9: Depth+Normal→RGB. Image-to-image translation results on the validation split of Taskonomy.

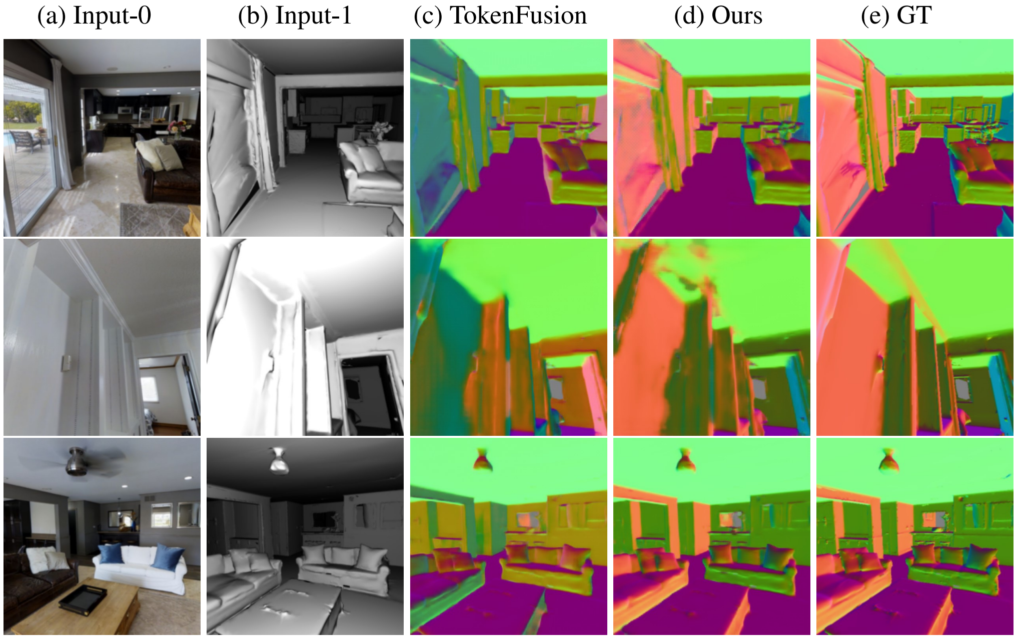


Figure 10: RGB+Shade→Normal. Image-to-image translation results on the validation split of Taskonomy.