Subset-Consistent Latent Reconstruction for Robust RGB-D Segmentation

Ekansh Mittal
Stanford University
450 Jane Stanford Way, Stanford, CA 94305
ekanshm@stanford.edu

Sarang Goel
Stanford University
450 Jane Stanford Way, Stanford, CA 94305
sarang28@stanford.edu

Abstract

Fusion of RGB and Depth sensors allows an accurate scene understanding and is crucial for applications in modern robotics and augmented reality. In practice, factors like poor lighting, motion blur, or even transparency result in severe sensor degradation that traditional fusion systems are not robust enough to handle. In this paper, we introduce an RGB-D semantic segmentation system based on latent reconstruction, which can reconstruct stable fused embeddings from corrupted or missing RGB and Depth data. Our system is based on SegFormer-B2 RGB-D fusion and introduces a latent restoration module. To train this module, we employed a teacher-student framework that teaches the latent restoration module to restore clean teacher model latents from corrupted student latents. To show the advantage of the proposed approach, we performed segmentation experiments on the NYUv2 dataset and demonstrated that the latent restoration module is able to improve the segmentation results on corrupted data, particularly on depth-only data. We see that although latent reconstruction-based semantic segmentation systems are still limited in the case of missing data, our framework is able to mitigate sensor degradation to help employ more reliable vision systems.

1. Introduction

1.1. Background and Motivation

In the context of modern robotics and Augmented Reality, the combination of RGB and depth is critical for understanding a scene. However, for systems to be deployed in the real world, they must be resistant to input degradation caused by environmental factors. Standard RGB systems typically fail with poor lighting, motion blur, and occlusions, among others [8]. On the other hand, depth fails in cases with transparent and reflective surfaces, among others. Since both systems have their unique problems and constraints, we can not expect a system that relies on both to work when the inputs of one have been degraded, or turned off entirely. This motivates the development of scene understanding systems and methods that can operate under missing/degraded sensor conditions.

Our goal is to develop a system capable of maintaining stable and accurate scene understanding regardless of sensor degradation. More specifically, we attempt to train an RGB-D model that produces stable embeddings from RGB only, depth only, clean RGB + depth, and corrupted versions of these inputs. We evaluate the quality of the embeddings on a semantic segmentation task [12]. Prior work has primarily focused on transformer-based methods to fuse RGB with other input modalities. However, these approaches heavily prioritize the fusion mechanism as opposed to focusing on robustness under modality dropout/corruption [5] and do not generalize to cases where modalities are missing or corrupted. Thus, an approach that dynamically aligns such sensor readings into a unified semantic space is critical to deploy reliable vision systems in production.

1.2. Proposed Approach

The goal of the multi-modal robustness we design is to approximate the presence of a clean RGB-D Teacher. Our algorithm faces the challenge of the RGB images and/or the depth maps being absent or corrupted. First, we train a SegFormer-B2 RGB-D Fusion Model. Then, a residual restoration module is appended to the model and tasked with correcting condition-specific latents, which are passed to the segmentation decoder to predict a semantic mask. Our loss design steers the model, using the teacher latent anchor, towards outputs that are semantically similar to the degraded inputs. This is done in order to enforce a certain level of semantic cohesion on the degraded outputs. Unlike the traditional deterministic fusion models, which struggle to accomplish this, we believe that a teacher-anchored, latent reconstruction of the model can provide a valid semantic structure because the model can be guided to fill in the gaps and, even with strong signal degradation, generate the missing semantic content.

2. Related Work

Scene understanding has advanced with multi-modal frameworks and the use of RGB and depth sensors. Attempts to integrate features can be classified into three groups: deterministic multi-modal fusion, distillation-based missing modality imputation, and parameter-efficient decoupled adaptation.

2.1. Deterministic Multi-Modal Fusion

In earlier studies, the main focus was on merging the input modalities through early or intermediate feature fusion. The alignment of the modalities was performed through mostly static strategies, with little or no dynamic feature learning. For instance, in Zhang et al. [9], the Semantic Guidance Fusion Network mitigated the early fusion rigidity by merging RGB and depth data through the use of intermediate learnable convolutions. Also, Zhang and Xie [10] advanced the segmentation of RGB and depth data by proposing MIPANet, which utilized multi-modal interaction and pooling attention to reduce the initial segmentation gap between the two input modalities. More recently, Brödermann et al. [1] developed CAFuser, a more advanced, condition-based method, which creates condition tokens to dynamically adjust the weights of different sensor modalities in adverse environmental conditions. These approaches do achieve high accuracy in the presence of full unimpaired sensor data, but their main drawback is the deterministic nature of their structure. In the case of a corrupted or missing modality, the networks collapse, since they were designed to expect complete unimpaired multimodal tensors at the time of inference.

2.2. Distillation and Masked Representation Learning

To deal with the risk of a complete dropout of a sensor, a separate class of methods uses self-teaching and cross-modal distillation. These do not rigidly fuse inputs. Instead, they focus on training the model to 'hallucinate' the missing privileged knowledge. For example, in Giannone and Chidlovskii [2], standard input fusion was replaced with a common deep representation allowing one available modality to reconstruct the other in missing-modality cases. Inspired by this, Zhao et al. [12] developed MaskMentor, where sensor failure was simulated by random modality and patch masking, both applied in a controlled way, and supervised by a complete-modality teacher. In a similar fashion, Wei et al. [6] developed a one-stage modality distillation framework, where the privileged knowledge was captured and restored through a multi-task learning approach. These frameworks are useful as they force the models to maintain the structural integrity of the latent space. But, they depend on several highly complex and expensive steps, and on frozen teachers.

2.3. Parameter-Efficient & Decoupled Adaptation

The most advanced methods are moving away from inflexible fusion and aggressive distillation and adapting towards the decoupling of representations and parameter-efficient methods. Reza et al. [4] showed that the core model of a pre-trained multimodal network can remain intact while low-rank adaptable layers can be added to compensate for the absence of modalities. To decouple representations, Wei et al. [7] suggested that, instead of placing modality combinations at fixed points in the latent space, they can be modeled as a probabilistic distribution, and thus, avoid the rigid constraints of intra-class relationships. In dual-modal systems, Hao et al. [3] introduced the Conditional Dropout learning approach, which maintains the integrity of the system for both full and partial data situations. To deal specifically with the discrepancies of different modalities, Zhao et al. [11] proposed a method of "bridging-then-fusing" in which alignment of modalities occurs before fusion.

Of these, we think that approaches based on conditional dropout and parameter-efficient adapters [3, 4] are the most refined and state of the art, since they solve the dynamic problem of modality dropout, and do not require a feature alignment or the use of a disconnected sub-network. By focusing on a teacher-driven latent reconstruction, our proposed method aims to build on these adaptive strategies, which retain structural integrity.

3. Methods

Our engineering goal is to develop a system that generates consistent latent representations on incomplete and degraded sensor inputs, as evaluated on semantic segmentation. We used the NYUv2 dataset to train our model, which has RGB images alongside aligned depth maps and pixel level semantic label maps. We train a single model capable of handling each sensor subset/corruption pair, conditioned on a modality-availability mask. We use a pretrained SegFormer MiT-B2 backbone for RGB-D feature extraction, alongside our main contribution, which is a latent restoration module that reconstruct clean RGB-D semantic features from missing or corrupted modality inputs.

3.1. Dataset and Preprocessing

As stated before, our project uses the NYUv2 labeled RGB-D dataset, which has predefined training set had 795 samples and our validation/test set had 654 examples. The labels consisted of 40 different semantic classes defined in the dataset. For preprocessing, all images, depth maps, and labels were resized to 240×320. We applied bilinear resizing to images and depth maps and nearest neighbor interpolation to resize labels. Further, RGB images were scaled to $[0, 1]$ and normalized using ImageNet statis-

tics with mean $[0.485, 0.456, 0.406]$ and standard deviation $[0.229, 0.224, 0.225]$. The depth images were also scaled to $[0, 1]$ and normalized using mean $[0.5, 0.5, 0.5]$ and standard deviation $[0.5, 0.5, 0.5]$. The depth values were repeated three times for SegFormer encoding.

3.2. Robustness Conditions

We evaluated our models under five input conditions:

- **clean_rgbd**: both RGB and depth are available
- **rgb_only**: depth is removed by zeroing the depth tensor
- **depth_only**: RGB is removed by zeroing the RGB tensor
- **rgb_corrupt**: RGB is degraded while depth remains clean
- **depth_corrupt**: depth is degraded while RGB remains clean

We ensured that the model was conditioned on which sensors were available by adding a two-dimensional modality availability mask. We used four different corruption methods for RGB, namely low-light scaling, gaussian noise, average blur, and random occlusion patches. For depth corruption, we used mixed depth noise, random holes, and invalid-pixel masking. The validation images themselves are held out from training, but the corruption families are not held out, with the model being trained and evaluated using the same corruption protocol for RGB and depth degradation. Thus, these experiments measures generalization to held-out NYUv2 scenes under controlled sensor degradation, rather than generalization to entirely unseen corruption types. Mixed RGB corruption applies low-light scaling, Gaussian noise with standard deviation 0.08, a 5×5 average blur, and two random zero-valued occlusion patches. Mixed depth corruption applies Gaussian noise with standard deviation 0.06, three random zero-valued hole patches, and an invalid-pixel mask with drop probability 0.16. All corrupted values are clipped to $[0, 1]$.

3.3. SegFormer-B2 RGB-D Fusion Model

Our approach was built off the pretrained SegFormer MiT-B2 from nvidia. We finetuned two parallel encoders end to end for RGB and depth respectively. For the fusion process, we used the final four hidden-state feature maps, and projected each to a shared 256 latent using 1×1 convolution, group normalization, and GELU activation. We apply the modality mask prior to fusion to ensure that unavailable modalities contribute zero features. Our lightweight gated fusion block was inspired by cross-modal rectification

Given projected RGB features r_s and depth features d_s at scale s, we compute two gates:

$$g_{\text{rgb}} = \sigma(W_{\text{rgb}}\,[r_s, d_s])\,, \tag{1}$$

$$g_{\text{depth}} = \sigma(W_{\text{depth}}\,[r_s, d_s])\,, \tag{2}$$

where $\sigma(\cdot)$ denotes the sigmoid function. The rectified features are

$$r'_s = r_s + g_{\text{rgb}} \odot d_s, \tag{3}$$

$$d'_s = d_s + g_{\text{depth}} \odot r_s, \tag{4}$$

These rectified features are then concatenated and passed through convolutional fusion layers.

The decoder then upsamples the four fused feature maps to the highest feature resolution, concatenates them, and predicts 40-class semantic logits, which are then bilinearly upsampled back to the original input resolution. We train the model using semantic cross-entropy, which is defined as:

$$\mathcal{L}_{\text{fusion}} = \text{CE}(\hat{y}, y). \tag{5}$$

This served as our baseline for semantic segmentation on RGB-D fusion.

During training the SegFormer fusion model samples a single input condition per batch on the following distribution: clean_rgbd (45%), rgb_only (15%), depth_only (15%), rgb_corrupt (12.5%), and depth_corrupt (12.5%). We prioritized clean_rgbd to ensures that the model can learn robust embeddings of missing and corrupted sensors while still perform well on clean inputs.

3.4. Latent Restoration Model

After finetuning the SegFormer-B2 fusion model, we add a latent restoration module, which we train to reconstruct the latents of clean inputs from degraded inputs. We use a student-teacher framework, where the teacher is a frozen checkpoint of the best SegFormer fusion model. We initialize the student model with the same weights. For each sample, the teacher produces a clean semantic `z_clean` and semantic segmentation logits from the clean RGB-D input. The student then takes the degraded input and produces the condition specific latent, denoted `z_cond`. The difference between the student and the teacher is that the student has an additional residual restoration network that predicts a correction, which is computed as:

$$\delta = R(z_{\text{cond}}, m), \tag{6}$$

$$z_{\text{restored}} = z_{\text{cond}} + \delta, \tag{7}$$

where m is the modality mask. The restored latent is then decoded by the SegFormer fusion decoder to produce the final segmentation logits.

For each training batch, the teacher runs its forward pass only on clean RGB-D input, which is used to supervise the student on two degraded views from rgb_only, depth_only, rgb_corrupt, and depth_corrupt, which are independently sampled with weights $[0.30,\ 0.30,\ 0.20,\ 0.20]$. We also supervise the students on the clean RGB-D input, to ensure that it stays strong on clean inputs. We made the missing modality case slightly more frequent than the corruption case while training the student because they are the more severe failure mode and produced the largest baseline performance drops.

For latent restoration our loss formulation consists of 1) segmentation loss on clean inputs, 2) segmentation loss on degraded inputs, 3) latent restoration loss compared to the teacher, and 4) optional teacher consistency loss on segmentation output distribution.

The latent restoration loss encourages the restored latent to match the frozen teacher latent using Smooth L1 distance. This is defined as:

$$\mathcal{L}_{\text{rest}} = \text{SmoothL1}(z_{\text{restored}}, z_{\text{clean}}). \tag{8}$$

The full no-consistency loss is

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + 0.25\,\mathcal{L}_{\text{rest}} + 1.0\,\mathcal{L}_{\text{student\_ce}} \tag{9}$$

where $\mathcal{L}_{\text{seg}}$ is the segmentation loss on the clean inputs and $\mathcal{L}_{\text{student\_ce}}$ is the segmentation loss on the degraded inputs.

The full-consistency variant additionally uses a latent MSE loss and a segmentation-level KL distillation loss from the teacher:

$$\mathcal{L}_{\text{latent}} = \text{MSE}(z_{\text{restored}}, z_{\text{clean}}), \tag{10}$$

$$\mathcal{L}_{\text{pred}} = \text{KL}\Big(\text{softmax}\Big(\frac{y_{\text{teacher}}}{T}\Big), \text{softmax}\Big(\frac{y_{\text{student}}}{T}\Big)\Big), \tag{11}$$

with temperature $T = 2.0$. The full objective is

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{seg}} + 0.25\,\mathcal{L}_{\text{rest}} + 1.0\,\mathcal{L}_{\text{student\_ce}} \\ + \lambda_{\text{latent}}\,\mathcal{L}_{\text{latent}} + \lambda_{\text{pred}}\,\mathcal{L}_{\text{pred}},\end{aligned} \tag{12}$$

where $\lambda_{\text{latent}} = 0.05$ and $\lambda_{\text{pred}} = 0.025$. The reason why we incorporate the MSE into our model apart from the Smooth L1 loss is due to the fact that the two have slightly different effects on the student network. Smooth L1 is the main restoration loss because it is more robust against large latent error when it comes to the first iterations of training where the student's features are quite far away from the teacher's features. MSE is an auxiliary loss which imposes a stronger constraint for smaller deviations of the restored latent representation from the teacher's more strongly once the restored representation is closer to the clean target. Because this stricter constraint can over-regularize the student, we apply it only in the consistency model and anneal its weight during training. These two additional consistency weights were kept at full strength for the first 20% of epochs, then decayed linearly until 80% of training, until they reached 10% of their initial values, which let the teacher provide stronger guidance to the student at the start while giving the student more freedom later on.

3.5. Optimization

We used the AdamW optimizer to train the model for 100 epochs, using learning rate 6e-5, weight decay 0.01, cosine learning-rate decay, and minimum learning rate 1e-6. We used a batch size of 4 to train the initial fusion model but then used a batch size of 2 for the latent restoration model since each training step required a forward pass of the frozen teacher and multiple forward passes of the student model.

After each epoch, we validated on all 5 conditions and chose the final model as the one with the highest average robust mIoU, which is the mIoU over the four non-clean conditions. We report average robust mIoU, the mIoU on each individual condition, and the max mIoU drop, which is defined as the largest decrease from clean RGB-D mIoU to any robustness condition. In addition to the mIoU based metrics, we the pixel accuracy on clean inputs and averaged over non-clean conditions, and the mean class average on clean inputs and averaged over non-clean conditions.

4. Experiments, Results, and Discussion

We perform robust RGB-D semantic segmentation on NYUv2 under the following five input scenarios: clean RGB-D, RGB-only, depth-only, corrupted RGB, and corrupted depth. The clean RGB-D scenario tests the conventional semantic segmentation accuracy, where all sensors are providing measurements. On the other hand, the RGB-only and depth-only data scenarios test robustness against missing modalities, whereas the corrupted RGB and corrupted depth data scenarios test robustness against corrupted sensory data. The central comparison is made among the following three SegFormer-B2 architectures: a pretrained RGB-D fusion baseline, the same model with latent restoration, and latent restoration along with additional teacher consistency losses. Furthermore, our own custom RGB-D transformer fusion model is also included as it served as our initial reference point and motivated the use of a stronger pretrained backbone.

4.1. Evaluation Metrics

Our primary evaluation metric is mean intersection-over-union (mIoU),which is common to semantic segmentation because it ensures balance between all categories instead of favoring one category that covers a larger area, like walls and floors. Given a category k, define true positives as TP_k,

false positives as FP_k, and false negatives as FN_k. The class IoU and mIoU are

$$IoU_k = \frac{TP_k}{TP_k + FP_k + FN_k}, \tag{13}$$

$$mIoU = \frac{1}{K}\sum_{k=1}^{K} IoU_k. \tag{14}$$

where K is the number of semantic classes. We also report pixel accuracy and mean class accuracy. Pixel accuracy measures the fraction of correctly classified pixels:

$$PixelAcc = \frac{\sum_k TP_k}{\sum_k (TP_k + FN_k)}. \tag{15}$$

Mean class accuracy averages per-class recall:

$$MeanClassAcc = \frac{1}{K}\sum_{k=1}^{K} \frac{TP_k}{TP_k + FN_k}. \tag{16}$$

Because our goal is to ensure robustness for missing or corrupted sensors, we additionally report average robust mIoU, defined as the mean mIoU over the four non-clean conditions:

$$AvgRobust = \frac{1}{4}(mIoU_{\text{RGB-only}} + mIoU_{\text{Depth-only}} + mIoU_{\text{RGB-corrupt}} + mIoU_{\text{Depth-corrupt}}). \tag{17}$$

We also report the maximum robust drop, which is the largest decrease from clean RGB-D mIoU to any degraded condition. This captures the worst-case sensitivity of a model to missing or corrupted input.

4.2. Experimental Setup and Hyperparameters

All the models were trained and evaluated on the official NYUv2 train/validation split. K-fold cross-validation was not done because there is an official validation split on NYUv2 and training multiple variants of the SegFormer-B2 architecture with restoration modules is very computationally expensive. The validation split was used both for model selection and performing the robustness analysis. The checkpoints for the SegFormer experiments were chosen based on the average robust mIoU and not the clean mIoU, as the goal of the project is not only maximizing the clean segmentation accuracy but also maintaining the accuracy when there is a missing or corrupted sensor.

The SegFormer-B2 experiments utilized AdamW optimizer with a learning rate of 6×10^{-5}, weight decay of 0.01, cosine learning-rate decay schedule, and mixed-precision training. This learning rate was selected as it was sufficiently low for fine-tuning the transformer backbone on the pretrained architecture while also high enough to allow fusion and restoration modules to adjust to RGB-D data input. The SegFormer-B2 fusion model used a mini-batch size of 4 while the restoration models used a mini-batch size of 2 since the latter incorporate both student model and frozen clean RGB-D teacher pass, which consumes more GPU memory. All SegFormer-B2 models were trained for 100 epochs during which we randomly sample both clean and noisy input conditions so that the model was able to see both clean RGB-D inputs and the other missing/corrupted inputs. For the consistency model, the teacher consistency losses were annealed during training to prevent excessively over-constraining the student in the beginning of optimization process.

4.3. Quantitative Results

Table 1 presents an overview of mIoU performance across all evaluation configurations. The custom transformer fusion baseline achieves low scores relative to SegFormer-B2 architecture, obtaining only 0.2508 clean mIoU and 0.1124 average robust mIoU. Such a low score indicates the difficulty of learning the representations of RGB-D semantics from scratch on the NYUv2 dataset. Instead, using the pre-trained SegFormer-B2 model in the place of custom encoder results in higher scores of 0.4648 clean mIoU and 0.3996 average robust mIoU.

The baseline SegFormer-B2 fusion model is relatively robust, with a clean mIoU is 0.4648, while RGB-only, RGB-corrupt, and depth-corrupt remain above 0.40 mIoU. Depth-only performs the worst with a score of 0.3372. This indicates that RGB has greater discriminative semantic value than depth for many NYUv2 classes. Depth is excellent at indicating surfaces, room layout, and large furniture but not great for separating visually identical semantic classes.

The addition of latent restoration leads to improved results in almost all conditions. Without consistency losses, average robust mIoU increases from 0.3996 to 0.4076, a marginal gain of 0.0080. The greatest improvement occurs in the depth-only scenario with an increase in mIoU from 0.3372 to 0.3553. This suggests that the restoration module is most useful when the model has to infer a clean joint representation from a single weaker modality. The max robust drop also decreases from 0.1276 to 0.1124, meaning restoration reduces the worst-case failure gap.

Adding consistency losses achieves the best overall model. In comparison with the fusion model SegFormer-B2 used as the baseline, the entire model incorporating the consistency losses performs better with improvements in clean mIoU by 0.0063, RGB-only mIoU by 0.0084, RGB corrupted mIoU by 0.0127, and depth corrupted mIoU by 0.0055. As for the average robust mIoU, it rises from 0.3996 to 0.4101. In contrast to the restoration without consistency, the model outperforms in terms of average robust mIoU by 0.0025 and all conditions apart from depth-only.

The mIoU improvements also manifest themselves

Table 1. Condition-specific mIoU and robustness summary.

Model	Clean	RGB Only	Depth Only	RGB Corr.	Depth Corr.	Avg. All	Avg. Robust	Max Drop
Custom Transformer Fusion	0.2508	0.1011	0.0830	0.1588	0.1065	0.1400	0.1124	0.1678
SegFormer-B2 Fusion	0.4648	0.4241	0.3372	0.4094	0.4275	0.4126	0.3996	0.1276
SegFormer-B2 + Latent Restoration	0.4676	0.4287	**0.3553**	0.4173	0.4293	0.4196	0.4076	**0.1124**
SegFormer-B2 + Restoration + Consistency	**0.4711**	**0.4325**	0.3528	**0.4221**	**0.4330**	**0.4223**	**0.4101**	0.1183

Table 2. Clean and robust accuracy metrics.

Model	Clean Acc.	Clean MCA	Robust Acc.	Robust MCA
Custom Transformer Fusion	0.6045	0.3469	0.4036	0.1893
SegFormer-B2 Fusion	0.7442	0.5857	0.6945	0.5218
SegFormer-B2 + Latent Restoration	**0.7504**	0.5890	**0.7035**	0.5303
SegFormer-B2 + Restoration + Consistency	0.7501	**0.5904**	0.7030	**0.5316**

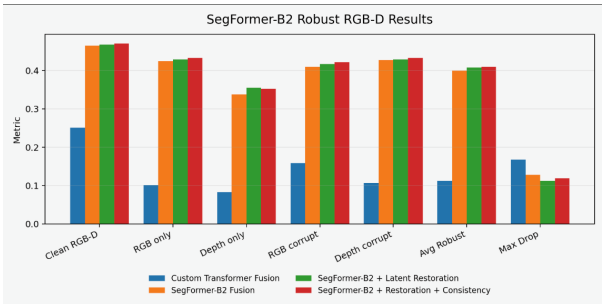

Figure 1. SegFormer-B2 Robust RGB-D Results

through improvements in pixel accuracy and average class accuracy as seen in Table 2. In terms of pixel and average class accuracy, pre-trained SegFormer-B2 backbone again provides the biggest improvement. Latent restoration increases robust pixel accuracy from 0.6945 to 0.7035 and robust mean class accuracy from 0.5218 to 0.5303. The robust mean class accuracy achieved by the consistency model is the highest of all three techniques at 0.5316, indicating that the consistency training method does not only increase accuracy on dominant large areas but slightly improves balanced class detection as well. The reason that pixel accuracy changes are smaller than mIoU changes is because NYUv2 contains large regions such as walls, floors, and furniture that dominate pixel counts.

The trends shown in Figure 1 are similar to those observed in Table 1. The SegFormer-B2 models show significantly higher scores than the custom transformer baseline and the restoration variants exhibit better robust performance consistently. Figure 2 shows that the validation results do not follow a monotonic increasing trend despite the decreasing nature of the loss curves during training. This suggests that the models can potentially start overfitting towards the sampled training distributions later during training. To counteract this, we utilized dropout, weight decay, randomizing the robustness condition, validation-based

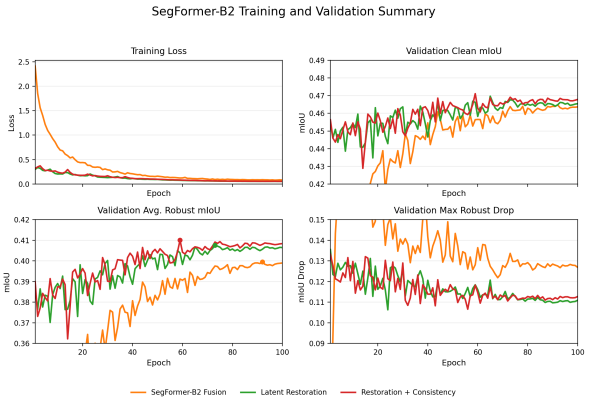

Figure 2. SegFormer-B2 Training and Validation Summary

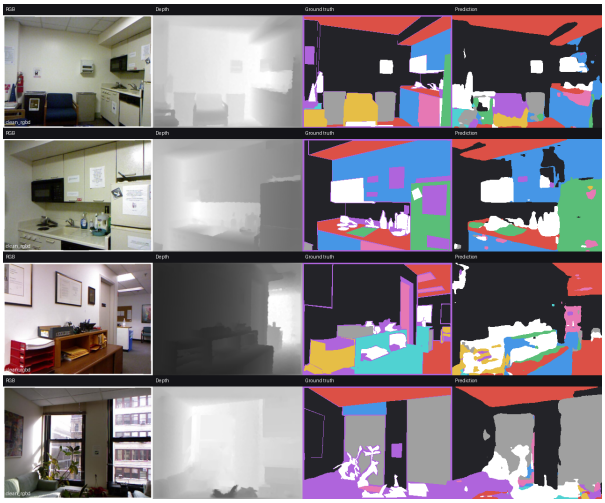

Figure 3. Clean RGB-D Segmentation Examples from SegFormer-B2 Restoration with Consistency Model

checkpointing, and, for the consistency model, annealed teacher losses.

4.4. Qualitative Analysis

Qualitative predictions for representative examples are shown in Figure 3. Under clean RGB-D and corruption-depth inputs, SegFormer-B2 recovers the basic shape and spatial structure of the rooms, along with large semantic regions. In general, restoration approaches recover coherent semantic regions better under degraded input conditions

than standard baselines, especially when one of the modalities is corrupted. Restoration aids in overcoming fragmented prediction due to noisy and dark parts of the images in RGB-corrupt scenes. When predicting only from depth alone, the model can recover large geometrical segments like walls, floors, and furniture but fails when segmentation requires appearance cues to distinguish between semantically different yet geometrically similar classes.

The main issue in qualitative failure is over-smoothing and semantic confusion in crowded indoor areas. Thin or small objects are often missing due to the low-resolution representation of the scene obtained after several downsampling steps within the encoder. Additionally, depth-only predictions also confuse classes with similar 3D structure. For instance, while the RGB appearance of furniture classes may differ, their depth profiles are very similar. This explains why depth-only remains the lowest mIoU condition even after restoration. These qualitative patterns match the quantitative results: latent restoration improves robustness, but it cannot fully reconstruct semantic information that is absent from the available sensor stream.

4.5. Discussion

We found that incorporating both latent restoration and consistency losses on top of the SegFormer-B2 backbone results in the highest performance, as demonstrated by it achieving the highest clean mIoU, RGB-only mIoU, RGB-corrupt mIoU, depth-corrupt mIoU, average all-condition mIoU, and average robust mIoU. This is not to discredit the no-consistency latent restoration model, which still achieeved the best performance on depth-only and the smallest worst-case drop in segmentation performance. These results show latent restoration is beneficial, but the gains were still modest in comparison to the gains provided by the pretrained SegFormer-b2 backbone over the custom transformer baseline.

The SegFormer-B2 model was crucial for our task since it provided strong semantic features before finetuning. The NYUv2 dataset is relatively small, which made it intractable to learn a transformer that can produce robust latent representations from scratch. The pretrained transformer gives the restoration module and stronger prior, and allows it to learn to move degraded semantic features toward the clean RGB-D teacher representation, which is why it helps the most under missing-modality and corruption settings.

It is important to note that the improvement from restoration was limited, since the baseline was already strong, leaving less room for improvement. Our lightweight restoration module also only takes one step, as opposed to a full iterative generative model like diffusion. In any case, some missing information is genuinely unrecoverable, which we saw with the poor performance of the depth only model. It was clear that without RGB, it is much more difficult to distinguish certain semantic classes. Thus, latent restoration serves to be a useful robustness mechanism, but is not a complete solution to missing-modality RGB-D segmentation.

5. Conclusion and Future Work

In this paper we demonstrate that a latent restoration approach can be used to maintain stable RGB-D semantic segmentation despite missing or degraded sensor inputs. We evaluated three main SegFormer-B2 architectures, starting from a pretrained fusion baseline, upon which we built a latent restoration module trained with and without teacher consistency losses. The use of a pretrained backbone was motivated by observing the weak performance of our custom transformer-based fusion model trained on NYUv2.

Our experiments on the NYUv2 data demonstrate that using the pre-trained SegFormer-B2 fusion model established a robust foundation for multi-modal segmentation. We built upon this strong baseline by adding a latent restoration module that improved segmentation performance across almost all degraded conditions, particularly on the depth-only setting, despite it being a weaker modality for segmentation. This shows that our latent restoration approach was effective in inferring a clean joint representation from a weaker modality. We further tested a latent restoration module trained with consistency losses, which further increased performance, yielding the best clean mIoU and average robust mIoU.

However, we found that there are still inherent limitations in missing modality imputation, as demonstrated by our qualitative and quantitative analyses. We found that depth was particularly difficult across all variations of the model, likely since depth maps fail to distinguish objects that share objects with similar 3D geometries. Without RGB data, depth alone may not contain enough evidence to inform semantic segmentation. Thus, our findings demonstrate that latent restoration can improve model robustness against sensor degradation, but it is not a complete solution for entirely missing modality streams.

Integrating temporal dynamics can potentially address these limitations, as consecutive frames in sequence could potentially provide additional context to help disambiguate objects. We also hope to investigate whether generative priors like diffusion or using Vision-Language Models (VLMS) to provide zero-shot semantic guidance during the reconstruction process can provide a more robust way to handle missing context. Eventually, to make these systems feasible for production deployment, we must investigate optimizing such architectures to achieve real-time inference latency on constrained edge computing hardware, ensuring reliable operation in complex, real-world environments.

6. Contributions and Acknowledgements

Sarang Goel implemented the custom transformer and SegFormer-B2 finetuning, and restoration module. Ekansh Mittal implemented the image processing pipeline and consistency loss. The final report was written jointly.

AI declaration: We utilized AI assistance for the purposes of brainstorming ideas for the overall project, understanding certain section from papers of related works and other methodologies that currently exist, debugging image pre-processing code (including how to create corruption masks that reflect real-world sensor failure), visualization code, and LaTeX errors, and other general coding errors. All architectural decisions were made and implemented by the authors without the assistance of AI.

References

[1] T. Brödermann, C. Sakaridis, Y. Fu, and L. Van Gool. Cafuser: Condition-aware multimodal fusion for robust semantic perception of driving scenes. *IEEE Robotics and Automation Letters*, 10:3134–3141, 2025.

[2] G. Giannone and B. Chidlovskii. Learning common representation from rgb and depth images. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 408–415. IEEE, 2019.

[3] S. Hao, C. Zhong, and H. Tang. Cola: Conditional dropout and language-driven robust dual-modal salient object detection. *arXiv preprint arXiv:2407.06780*, 2024.

[4] M. K. Reza, A. Prater-Bennette, and M. S. Asif. Robust multimodal learning with missing modalities via parameter-efficient adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47:742–754, 2025.

[5] J. Tan, X. Zheng, and Y. Liu. Towards robust multi-modal semantic segmentation with teacher-student framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

[6] S. Wei, Y. Luo, and C. Luo. One-stage modality distillation for incomplete multimodal learning. *arXiv preprint arXiv:2309.08204*, 2025.

[7] S. Wei, Y. Luo, Y. Wang, and C. Luo. Robust multimodal learning via representation decoupling. *arXiv preprint arXiv:2407.04458*, 2024.

[8] G. Xie, M. Li, S. Wu, Y. Liu, Z. Xie, B. Cao, and H. Liu. Hdcnet: A hybrid depth completion network for grasping transparent and reflective objects. *IEEE Robotics and Automation Letters*, 2025.

[9] P. Zhang, M. Chen, and M. Gao. Semantic guidance fusion network for cross-modal semantic segmentation. *Sensors*, 24(8):2473, 2024.

[10] S. Zhang and M. Xie. Optimizing rgb-d semantic segmentation through multi-modal interaction and pooling attention. *arXiv preprint arXiv:2311.11312*, 2023.

[11] S. Zhao, Y. Liu, Q. Jiao, Q. Zhang, and J. Han. Mitigating modality discrepancies for rgb-t semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 35:9380–9394, 2024.

[12] Z. Zhao, J. Li, L. Wang, Y. Wang, and H. Lu. Maskmentor: Unlocking the potential of masked self-teaching for missing modality rgb-d semantic segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1915–1923, 2024.