RobuMTL: Enhancing Multi-Task Learning Robustness Against Weather Conditions

Supplementary Materials

DMLS Performance

The performance of the DMLS is summarized in Table 1, where we evaluate the model across several key metrics: Accuracy, Precision, Recall, and F1-score. Accuracy reflects the model's ability to make correct predictions, providing a general measure of its overall effectiveness. Precision quantifies how many of the instances predicted as positive are actually correct, while recall measures how well the model identifies all true positive instances. The F1-score, calculated as the harmonic mean of precision and recall, balances these two metrics to provide a comprehensive assessment of the model's performance. The Receiver Operating Characteristic (ROC) curve illustrates the capability of DMLS in discriminate between perturbation classes as shown in Figure 1. The DMLS provides good classification performance on rain, noise, blur, and fog classes. It may struggle slightly to differentiate between snow and clean images, as they share some common properties, and some images contain small snowflake patterns that lead the classifier to mistakenly classify them as clean. However, the 6% difference in accuracy does not significantly impact the performance of routing to select the top-K experts. This is because the scores are generated based on the total voting over the batch of images. Even if the classifier misclassifies one of the images, the overall voting across the batch will mitigate this issue achieving 99.9% accuracy.

Table 1. DMLS performance metrics.

Metric	Value
Accuracy (%)	94.44
Precision (%)	93.36
Recall (%)	92.58
F1-score (%)	92.65
No. Parameters	48,742

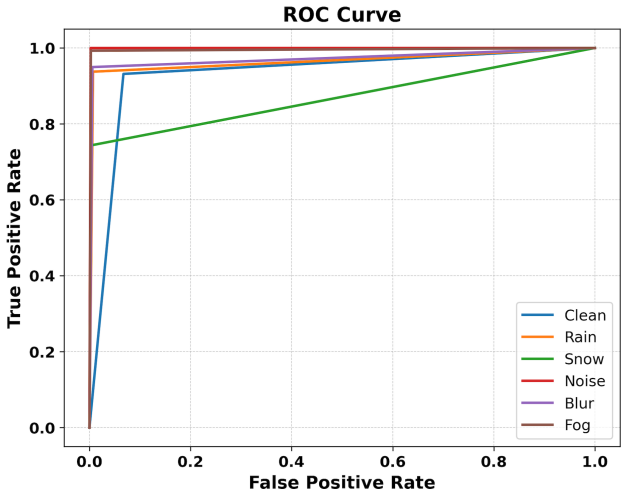


Figure 1. The DMLS performance ROC curve in classification of perturbations.

LoRA Ranks Performance on Perturbations

From the shown Figure 2, the LoRA configuration r[64,64,64,64] achieves the best overall performance on the Human Parts segmentation task under clean conditions, indicating that keeping a consistently strong representational capacity across all encoder stages benefits high-level semantic recognition. The r[8,16,32,64] configuration performs better under perturbations, especially noise and blur, but shows a noticeable drop in clean performance due to its lower capacity in early layers. The r[16,32,64,128] setup provides a balanced alternative, ranking third overall, and shows competitive robustness under certain weather degradations, but still does not reach the same clean-data peak as uniform higher ranks.

These trends reveal that Human Parts task relies heavily on global structural context and fine-grained semantic cues, which become more stable when the model maintains sufficient rank in shallow and mid-level stages (as in r[64,64,64,64]). Meanwhile, hierarchical growth schemes like r[8,16,32,64] and r[16,32,64,128] introduce useful reg-

ularization against high-frequency perturbations—by constraining early representations and allowing more flexibility deeper—but this comes at the cost of reduced clean-scene accuracy for a task that demands strong semantic detail at all feature scales. For the Normals task, the r[64,64,64,64] configuration produces the lowest error in clean conditions, but it is less robust against perturbations, particularly noise and blur, where its error noticeably increases. In contrast, r[16,32,64,128] provides the best overall trade-off between clean accuracy and robustness, achieving strong performance across multiple adverse conditions including snow, fog, and blur. Meanwhile, the r[8,16,32,64] hierarchy demonstrates high robustness against rain, snow, and noise, benefiting from stronger suppression of high-frequency corruption in the earliest layers. These trends reflect the sensitivity of surface normal estimation to low-level geometric consistency. Uniform high ranks allow the network to preserve rich detail under clean conditions, but the same flexibility can lead to over-amplifying noise when early-stage representations are not sufficiently constrained. Hierarchically increasing ranks (as in r[16,32,64,128]) strike a favorable balance by limiting noise propagation in shallow layers while expanding representational capacity deeper in the network where shape and orientation information is consolidated. Additionally, the strong noise robustness of r[8,16,32,64] suggests that aggressive early compression acts like a built-in denoising filter, preventing corrupted features from degrading later geometric predictions.

For the saliency estimation task, the r[32,32,32,32] configuration delivers the highest accuracy under clean conditions and ranks among the best for several perturbations, but it shows a notable drop in robustness when strong noise degradations are introduced. Both r[8,16,32,64] and r[16,32,64,128] follow closely behind, maintaining stronger resilience against most perturbations, especially those that distort structural boundaries. In contrast, rank settings such as r[8,16,32,64], r[64,32,16,8], and r[128,64,32,16] demonstrate lower clean-data accuracy, indicating that either overly limited early-layer capacity or reversed/imbalanced rank hierarchies can weaken saliency recognition in ideal conditions.

This behavior aligns with the nature of saliency detection, which requires both mid-level semantic consistency and preservation of fine spatial cues. A uniform moderate rank like r[32,32,32,32] offers the right amount of global context extraction and detail representation, explaining its superior clean performance. However, the improved robustness of r[8,16,32,64] and r[16,32,64,128] suggests that progressively increasing ranks across the network help buffer against corruption, where early compression filters out noise and deeper layers retain enough flexibility to reconstruct meaningful foreground attention cues.

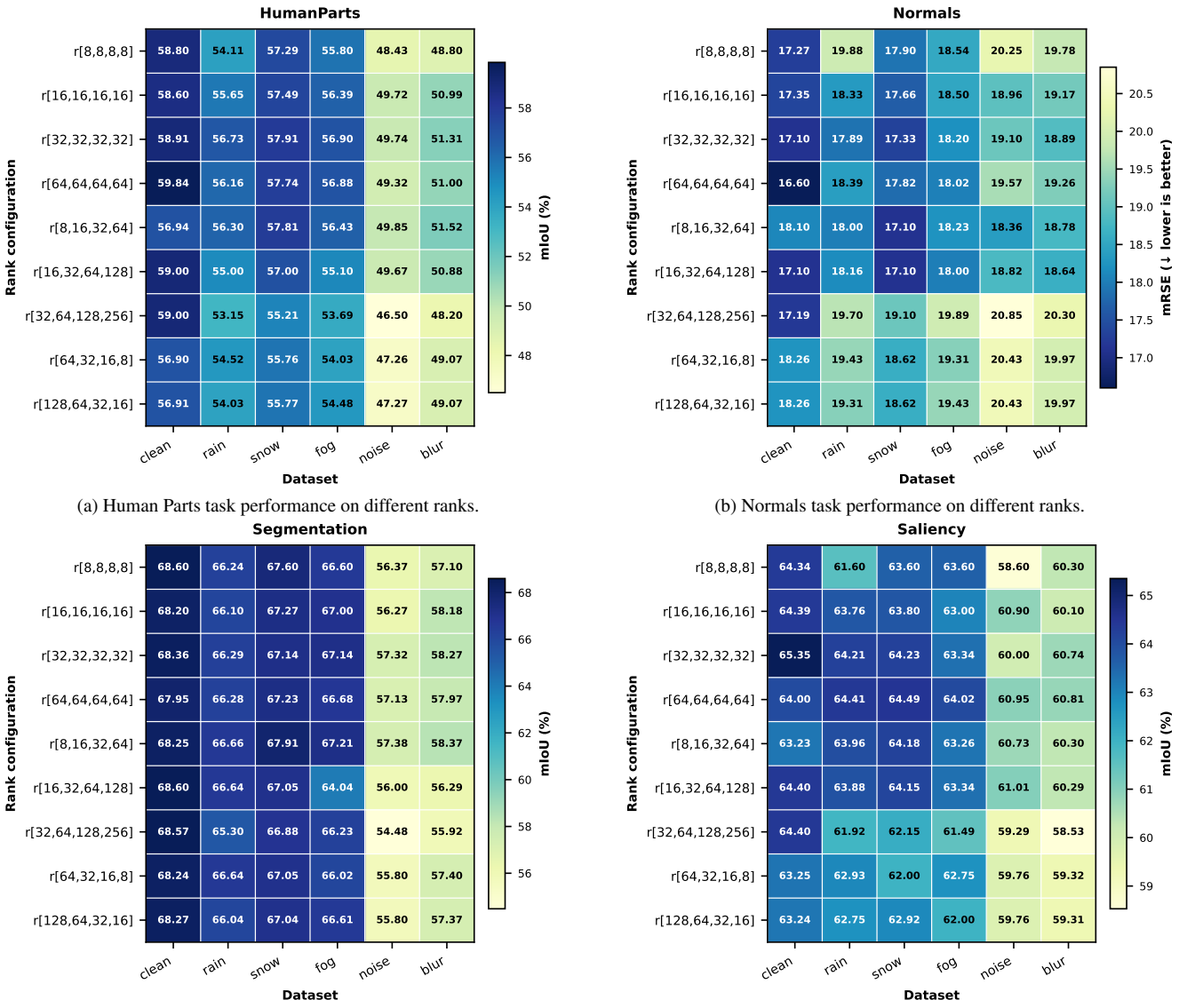
For the segmentation task, the r[8,16,32,64] configuration achieves the highest robustness across all perturbation types, showing strong accuracy even under challenging degradation, with only a slight decrease in clean performance. The r[8,8,8,8] configuration follows as the next most consistent choice under corruption, demonstrating that a uniformly low-rank design can effectively filter noise, though it lacks the deeper representational strength needed for peak accuracy in clean scenes. Meanwhile, r[16,32,64,128] delivers the best clean-data performance and excels specifically under snow, benefiting from increased rank capacity in deeper layers where semantic boundaries are refined.

This pattern highlights a key insight: segmentation models require a balance between noise suppression in early feature extraction and semantic expressiveness in deeper layers. Increasing ranks hierarchically (as in r[8,16,32,64]) allows the model to localize corruptions early while progressively restoring discriminative features, making it the most reliable choice across perturbations. In contrast, while higher ranks throughout the network (as in r[16,32,64,128]) can maximize clean-scene accuracy, they become more sensitive to corruption due to the higher freedom in feature adaptation, causing performance to deteriorate more noticeably under harsh perturbations.

Overall, our results consistently demonstrate that forward hierarchical LoRA rank configurations outperform uniform and reverse configurations in multi-task learning. In particular, the r[8,16,32,64] setup delivers the highest relative performance gains under perturbations across all tasks, effectively suppressing high-frequency noise in early layers and enabling deeper layers to recover semantic structure, although it exhibits slightly lower accuracy on clean data. Meanwhile, r[16,32,64,128] achieves the best balance between clean-scene performance and robustness, offering both strong baseline accuracy and stable degradation behavior across weather-related corruptions. These findings suggest that progressively increasing rank with network depth is a crucial design principle for enhancing MTL resilience: early-stage compression restricts noise propagation while deeper high-capacity representations preserve task-specific semantic information. In contrast, reverse hierarchical ranks perform consistently worse, indicating that allocating higher representational capacity to shallow layers is neither effective for semantics nor for robustness. Thus, forward hierarchical LoRA assignments, especially r[8,16,32,64] and r[16,32,64,128], provide the most reliable and scalable adaptation strategies for perturbation-aware multi-task learning

Frame Per Second (FPS) Evaluation

In our evaluation, the Frames Per Second (FPS) metric was calculated based on the average inference time measured over randomly sampled and combined images from all per-



(a) Human Parts task performance on different ranks.

(b) Normals task performance on different ranks.

(c) Semantic Segmentation task performance on different ranks.

(d) Saliency task performance on different ranks.

Figure 2. Breakdown of task performance across different LoRA rank configurations on clean and perturbed PASCAL datasets.

turbation datasets, following:

$$\text{FPS} = \frac{NB * BS}{T_{\text{seconds}}},$$

where NB is the number of processed batches, BS is the batch size, and T_{seconds} is the total time taken to process the frame. Unlike the original baseline, which reports FPS using batch size = 1, the model gives higher throughput using batched inference as well, as GPUs process multiple images more efficiently in parallel. To ensure fairness in comparison, we did all the evaluation on same GPU and same number CPU cores. Also, we treat each task-specific single-task model independently and report the minimum FPS among them. Importantly, the MTL model achieves faster overall inference than single-task deployment because it shares the encoder and common computations across tasks, reducing redundant processing. Although our proposed MTL variant achieves slightly lower FPS than the baseline MTL, this overhead is expected due to the additional modules (DMLS and MEFP), which increase robustness and task interaction at a modest cost in speed. RobuMTL(+) adds just 3.6 ms per image (17 ms total), with 0.5 ms for DMLS, 2.5 ms for LoRA aggregation, and 0.6 ms for injection, enabling robust inference with minimal overhead.

Model Performance Evaluation

We evaluate performance by averaging results across each perturbation type for both PASCAL and NYUD-v2. Unlike PASCAL, NYUD-v2 is smaller and includes edge and depth tasks, which make the MTL setup more sensitive to feature degradation. Although the same strategy is applied to both datasets, NYUD-v2 requires slightly higher ranks in the early layers due to its different data distribution and the need for stronger representation in edge and depth tasks.

Model Training

We explored multiple training paradigms and found that training noise- and blur-specific experts solely on their own data led to overfitting and degraded performance in tasks such as normals. In addition, applying standard MTL conflict-resolution techniques further worsened results, as noise amplification increased the RMSE for normals, depth, and edges.

We also experimented with an auxiliary consistency loss to align the model under perturbations with a clean teacher model. Initial observations showed up to a 1% improvement in some tasks, but this requires further investigation and analysis as part of future work.