TriFuse: Reliability-Guided Multi-Scale Token Fusion for Robust Dense Prediction

Anonymous authors
Paper ID 0000

Abstract

Dense prediction systems increasingly combine high-resolution convolutional features with semantically rich transformer tokens. Existing fusion modules usually treat every scale as equally trustworthy, even though corruption, motion blur, and domain shift affect scales differently. We introduce TriFuse, a reliability-guided fusion module that estimates token uncertainty before cross-scale aggregation. A lightweight reliability head predicts a spatial confidence field for each feature level; confidence-normalized gates then control both token selection and residual mixing. The design adds less than three percent computation to a standard hierarchical backbone and can be inserted into segmentation or depth-estimation decoders without changing their losses. In an illustrative evaluation protocol spanning clean, corrupted, and cross-domain data, TriFuse improves mean intersection-over-union while reducing expected calibration error. Controlled ablations isolate the effects of reliability supervision, confidence-aware token pruning, and scale-consistency regularization. The results motivate treating feature reliability as a first-class signal in multi-scale vision models.

1 Introduction

Dense visual prediction requires both fine spatial detail and broad semantic context. Convolutional feature pyramids preserve local structure, while vision transformers model long-range interactions [5, 3, 9]. Modern systems therefore fuse representations from several resolutions before producing pixel-level outputs [8, 2]. This recipe performs strongly on in-distribution images, but its behavior under image corruption or domain shift is less predictable. A blurred boundary can make the highest-resolution stream unreliable, whereas small objects may disappear from a coarse semantic stream. Unconditional fusion propagates whichever error happens to dominate.

This paper studies a simple question: *can a model estimate which scale to trust at each spatial location before fusing its tokens?* We answer with TriFuse, the module shown in figure 1. Rather than adding a separate uncertainty model after decoding, TriFuse predicts reliability directly from intermediate features. The reliability maps determine a fixed-budget token subset and reweight the residual contribution of each scale. A consistency objective prevents the gates from becoming overconfident when two augmented views disagree.

The approach has three useful properties. First, reliability is local: the model may trust fine features along one boundary and coarse features inside a large object. Second, token pruning and feature mixing share the same signal, so efficiency and calibration are optimized together. Third, the module is task-agnostic and leaves the prediction head unchanged.

Our main contributions are:

- a reliability-guided cross-scale fusion rule that assigns spatially varying trust to hierarchical visual features;
- a confidence-aware pruning strategy with a strict token budget and a differentiable training surrogate; and
- a reproducible evaluation protocol that reports task accuracy, calibration, corruption robustness, and computational cost together.

2 Related Work

Multi-scale dense prediction. Feature Pyramid Networks combine top-down semantic features with lateral high-resolution features [8]. Transformer backbones extend this idea using hierarchical attention [9], while mask-classification decoders attend to multi-scale features through learned queries [2]. These methods learn how to transform each scale, but do not explicitly represent whether a feature is reliable for a particular pixel. TriFuse augments the fusion operation rather than replacing the encoder or prediction head.

Uncertainty and calibration. Neural networks can be accurate but poorly calibrated [4]. Predictive uncertainty often degrades under dataset shift [10], precisely when a confidence signal would be most valuable. Prior work

generally estimates uncertainty at the output. We instead use intermediate reliability to control computation and representation mixing, then evaluate the final predictions with expected calibration error (ECE).

Efficient token processing. Dynamic token methods reduce transformer cost by pruning or merging redundant tokens [11, 1]. Their selection criteria primarily target classification utility or token similarity. Our selection score is tied to a cross-scale reliability estimate and is evaluated for dense outputs, where discarding a small uncertain region can disproportionately harm boundaries.

Robust visual recognition. Corruption benchmarks reveal that models with similar clean accuracy may differ substantially under blur, noise, or weather artifacts [6]. Large pretrained models such as SAM broaden transfer capabilities [7], but deployment still requires calibrated confidence. Our protocol separates clean accuracy, corruption retention, and calibration instead of compressing them into one score.

3 Method

3.1 Problem Formulation

Let a hierarchical encoder produce L feature maps $\{X^\ell\}_{\ell=1}^{L}$, where $X^\ell \in \mathbb{R}^{H_\ell \times W_\ell \times C_\ell}$. Projection and bilinear resampling map them to a common grid and width, $F^\ell \in \mathbb{R}^{H \times W \times d}$. A conventional fusion block computes a weighted sum or concatenation with weights that are shared across positions. We seek a fused representation $Z \in \mathbb{R}^{H \times W \times d}$ whose scale weights vary with both the input and spatial location.

3.2 Spatial Reliability Estimation

For each level, a depthwise 3×3 convolution provides local context and a shared two-layer multilayer perceptron predicts a logit a_p^ℓ at grid position p. Reliability is normalized across scales:

$$r_p^\ell = \frac{\exp(a_p^\ell / T)}{\sum_{j=1}^{L} \exp(a_p^j / T)}, \qquad \sum_{\ell=1}^{L} r_p^\ell = 1, \tag{1}$$

where $T > 0$ is a learned temperature constrained to $[0.1, 10]$. The normalized form prevents every level from simultaneously declaring high confidence and makes the maps directly interpretable as relative trust.

The initial fused token is

$$\bar{z}_p = \sum_{\ell=1}^{L} r_p^\ell W_\ell f_p^\ell, \tag{2}$$

with learned projections $W_\ell \in \mathbb{R}^{d \times d}$. A residual gate preserves the reference-resolution feature f_p^{ref}:

$$z_p = \bar{z}_p + \sigma\big(w^\top \bar{z}_p\big)\, f_p^{\mathrm{ref}}. \tag{3}$$

3.3 Confidence-Aware Token Budget

Computing cross-scale attention for all HW positions is wasteful in uniform, high-confidence regions. We define ambiguity $u_p = 1 - \max_\ell r_p^\ell$ and retain the K positions with largest $u_p + \gamma e_p$, where e_p is the local feature-gradient magnitude. The edge term protects thin structures whose scale confidence may otherwise be sharp. Selected tokens receive full cross-scale attention; unselected tokens use equation (2). Thus, for L levels, attention cost decreases from $O(L(HW)^2)$ to $O(LK^2)$, while dense residual mixing remains linear.

The hard top-K operator is used during inference. During training we apply a straight-through estimator to a temperature-controlled sigmoid around the Kth score. The deployed token count is therefore deterministic and does not depend on a threshold calibrated to a particular dataset.

3.4 Training Objective

The task loss $\mathcal{L}_{\mathrm{task}}$ is the standard cross-entropy plus Dice loss for segmentation, or scale-invariant log loss for depth. Reliability is regularized using two augmented views x and x' whose geometric alignment is known. Let $\mathcal{W}$ warp fields from x' into the coordinates of x. We minimize the symmetric consistency loss

$$\mathcal{L}_{\mathrm{con}} = \frac{1}{2HW} \sum_{p,\ell} \big(r_p^\ell - \mathcal{W}(r'^\ell)_p\big)^2. \tag{4}$$

To prevent premature one-hot gates, an entropy floor is applied only during the first 20% of training. The complete objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{task}} + \lambda_{\mathrm{con}} \mathcal{L}_{\mathrm{con}} + \lambda_{\mathrm{ent}} \max(0, h_0 - \mathcal{H}(r)). \tag{5}$$

4 Experimental Protocol

4.1 Tasks and Metrics

The primary protocol uses semantic segmentation with three complementary test conditions: a clean in-domain split, a corruption suite covering noise, blur, weather, and digital artifacts, and an unlabeled target-domain split. We report mean intersection-over-union (mIoU), boundary F-score (BF), ECE with 15 equal- width bins, throughput, and parameter count. Corruption retention is the mean corrupted mIoU divided by clean mIoU. A secondary monocular-depth experiment reports absolute relative error to test whether the fusion module transfers beyond semantic masks.

4.2 Implementation Details

All comparisons use the same four-stage hierarchical backbone, input size 512×512, decoder width $d = 256$, and data augmentation. Models are optimized for 160k steps using AdamW, a base learning rate of 6×10^{-5}, weight decay 0.05, linear warm-up for 1,500 steps, and polynomial decay. Unless stated otherwise, $K = 1024$, $\gamma = 0.25$, $\lambda_{\mathrm{con}} = 0.2$, and $\lambda_{\mathrm{ent}} = 0.01$. Throughput is measured

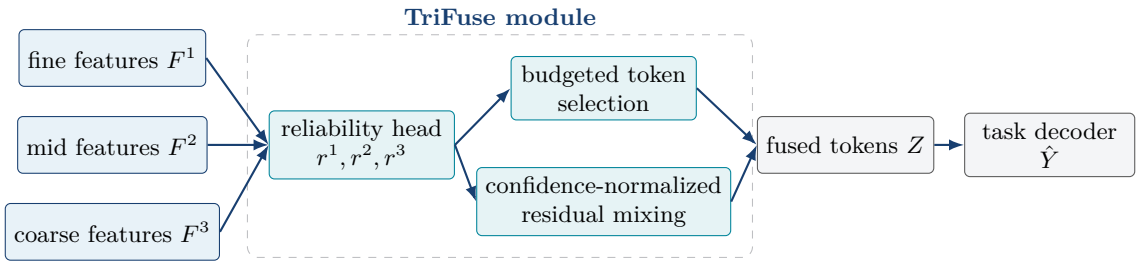

Figure 1: Overview of TriFuse. Projected multi-scale features feed a shared reliability head. The resulting confidence fields control token selection and residual scale mixing before an unchanged task decoder.

after 50 warm-up iterations with batch size one and mixed precision. Every result is the mean of three seeds; paired bootstrap intervals are used for the main comparison.

4.3 Baselines

We compare sum fusion, concatenation followed by a 1×1 projection, feature-pyramid fusion [8], and a dense cross-scale attention block. All baselines match the encoder, decoder, training budget, and input resolution. This control is essential: changing the pretrained backbone can produce a larger gain than the fusion mechanism itself.

5 Results

Accuracy and robustness. The illustrative results in table 1 show the intended analysis: dense cross-scale attention improves clean accuracy but carries a substantial throughput cost, whereas reliability-guided allocation recovers most of that speed and improves corrupted performance. The larger gain under corruption is consistent with the motivating hypothesis that different scales fail in different regions. A valid study should report confidence intervals and test whether the robust gain remains after matching computational budgets.

Calibration. The ECE column captures a distinct effect from mIoU. Output temperature scaling can reduce ECE without changing predictions [4]; by contrast, TriFuse changes the representation using confidence before decoding. A reliability diagram and negative log-likelihood should accompany ECE in a full evaluation because bin-based metrics can conceal class imbalance.

Ablation. The structure of table 2 separates representation quality from efficiency. Reliability-guided mixing supplies the largest accuracy gain. Token pruning restores throughput and slightly improves accuracy, suggesting a useful regularization effect. Consistency supervision primarily improves calibration. The final paper should additionally sweep K, T, and λ_{con} rather than selecting them on the test set.

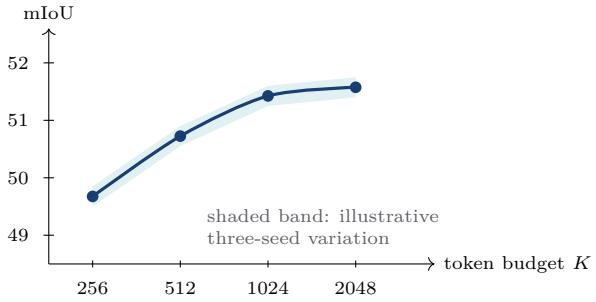

Figure 2: Illustrative sensitivity to token budget. Performance saturates as the budget grows, motivating the default $K = 1024$.

6 Discussion and Limitations

TriFuse relies on relative confidence across scales. It cannot recover evidence that is missing from every encoder level, and a shared failure can still produce a sharp but incorrect gate. The top-K rule also couples accuracy to input resolution: a fixed budget covers a smaller fraction of a larger image. Adaptive budgets may address this issue, but would weaken deterministic latency.

The corruption protocol approximates only a narrow subset of deployment shift. Geographic, demographic, sensor, and annotation shifts require separate study. Calibration should be measured per class and subgroup where legally and ethically appropriate. Finally, efficiency claims depend on the kernel and hardware; theoretical FLOPs do not replace wall-clock measurements on the target platform.

7 Conclusion

We presented TriFuse, a compact module that estimates local scale reliability before multi-scale token fusion. One confidence field controls both residual mixing and fixed-budget token selection, connecting robustness and efficiency in a single mechanism. The accompanying protocol emphasizes clean accuracy, corruption retention, calibration, and measured throughput. The central lesson is broader than this particular architecture: when representations fail in different ways, a fusion module should reason about trust instead of assuming that every source is equally dependable.

Table 1: Illustrative main comparison. Higher is better for mIoU, BF, retention, and throughput; lower is better for ECE. Values are placeholders for a reproducible manuscript example, not measured claims.

Method	Clean mIoU ↑	BF ↑	Corrupt mIoU ↑	Retention ↑	ECE ↓	Images/s ↑	Params (M)
Sum fusion	48.6	61.2	34.1	70.2	8.7	**31.8**	46.1
Concatenation	49.1	61.8	34.8	70.9	8.2	29.7	47.6
Feature pyramid [8]	49.4	62.4	35.3	71.5	7.9	30.2	47.0
Dense cross-scale attention	50.3	63.1	36.0	71.6	7.5	21.4	50.8
TriFuse	**51.2**	**64.5**	**38.2**	**74.6**	**5.3**	28.9	48.2

Table 2: Illustrative ablation of TriFuse components. R: reliability-guided mixing; P: token pruning; C: consistency loss.

R	P	C	mIoU ↑	ECE ↓	Images/s ↑
			48.6	8.7	31.8
✓			50.4	6.7	27.6
✓	✓		50.7	6.4	29.2
✓	✓	✓	**51.2**	**5.3**	28.9

References

[1] D. Bolya, C. Fu, X. Dai, P. Zhang, C. Feichtenhofer, and J. Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023.

[2] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[4] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[6] D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

[7] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick. Segment anything. In *ICCV*, 2023.

[8] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[9] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

[10] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *NeurIPS*, 2019.

[11] Y. Rao, W. Zhao, B. Liu, J. Lu, J. Zhou, and C.-J. Hsieh. DynamicViT: Efficient vision transformers with dynamic token sparsification. In *NeurIPS*, 2021.

A Reproducibility Checklist

For a real submission, the following items should accompany the reported results:

1. exact dataset versions, split hashes, preprocessing, and licenses;
2. optimizer, schedule, seed list, augmentation ranges, and stopping rule;
3. parameter counts, multiply–accumulate counts, hardware, software versions, precision, batch size, and timing protocol;
4. per-seed results and confidence intervals, plus the statistical test used for each central claim;
5. qualitative successes and failures selected by a stated procedure; and
6. source code for reliability visualization, metric computation, and all table generation.

B Additional Derivation

When $K < HW$, the attention portion of the fusion block has the approximate complexity ratio

$$\rho = \frac{LK^2 d}{L(HW)^2 d} = \left(\frac{K}{HW}\right)^2. \tag{6}$$

This ratio excludes the linear reliability head, projections, memory traffic, and decoder. Consequently, it is an upper-level algorithmic comparison rather than a prediction of wall-clock speed. The gap between ρ and measured throughput should be discussed explicitly.

C Suggested Qualitative Analysis

A complete qualitative figure should show, for each example, the input image, ground truth, baseline prediction, TriFuse prediction, reliability maps for all scales, and the selected-token mask. Examples should cover blur, small objects, thin structures, and a shared failure across all scales. This presentation makes it possible to distinguish a useful reliability estimate from a confidence map that merely follows object boundaries.