

SPAFNet: Semantic Prior-Guided Progressive Alignment and Gated Bidirectional Fusion for Unaligned RGB-T Salient Object Detection

Muye Chen

Northwestern University

Ming Wu

ming.wu1@kuleuven.be

KU Leuven <https://orcid.org/0000-0002-3582-4881>

Research Article

Keywords: RGB-T salient object detection, cross-modal alignment, dense correspondence estimation, semantic prior, confidence-aware fusion

Posted Date: October 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11440311/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

SPAFNet: Semantic Prior-Guided Progressive Alignment and Gated Bidirectional Fusion for Unaligned RGB-T Salient Object Detection

Muye Chen¹ Ming Wu^{2,3,*}

¹Department of Computer Science, Northwestern University,
Evanston, IL 60208, USA

²Department of Computer Science, KU Leuven, 3001 Leuven, Belgium

³Department of Mechanical Engineering, KU Leuven, 3001 Leuven, Belgium

*Correspondence: ming.wu1@kuleuven.be

Abstract

Unaligned RGB and thermal images require spatially tolerant fusion when parallax, sensor offset and modality-specific artifacts coexist. SPAFNet combines four-scale dense feature sampling with a shared semantic prior and gated bidirectional convolutional fusion. Sampling displacement is estimated from inherited coarse flow, local matching expectations and learned residuals; a saliency-supervised gate modulates thermal features without direct correspondence calibration. The supplied evaluation summary describes two naturally unaligned, three synthetically perturbed and three aligned datasets; the complete results remain to be documented in this draft. On UVT20K, the comparison table lists maximum F_m values of 0.887 for SPAFNet and 0.822 for PCNet. These unreplicated point estimates do not establish physically correct registration, statistical significance or performance on unseen imaging systems.

Keywords: RGB-T salient object detection; cross-modal alignment; dense correspondence estimation; semantic prior; confidence-aware fusion.

1 Introduction

RGB and thermal images provide different observations of a scene. Visible images describe appearance, texture and color, whereas thermal images describe emitted radiation in the sensor’s operating band. Their combination can support salient object detection when one observation is less informative. This complementarity depends on how evidence is associated across images. A thermal response at a given pixel need not describe the same scene point as the RGB pixel at the same location.

For unaligned image pairs, parallax, sensor offsets and unequal fields of view complicate direct feature fusion. A displacement that compensates for one region may be unsuitable elsewhere. Occluded or missing content also cannot be recovered simply by choosing another sampling position. A useful detector must therefore address where to obtain potentially corresponding features and how strongly to use those features.

SPAFNet is organized around these two questions. Its supplied design combines a shared semantic prior, progressive sampling over four feature scales, and gated bidirectional convolutional fusion. Coarse displacement is inherited at finer scales, local matching expectations supply nearby correspondence evidence, and learned residuals refine the sampling locations. A saliency-supervised gate controls the thermal contribution. Sampling and modulation have distinct roles even when both are learned for the detection task.

This draft develops the method-level description supplied by the authors and preserves the reported UVT20K comparison. It does not document a reproduced implementation or a completed independent evaluation. The abstract states coverage of eight datasets, but the complete tables, protocols and prediction files were not available for this draft. The mathematical notation below is schematic; implementation-specific choices remain to be documented.

2 Related work

Alignment-free RGB-T detection has already been studied through semantic guidance and cross-modal correlation. SACNet associates thermal evidence with RGB features through asymmetric correlation and feature sampling [1]. PCNet combines semantic guidance for homography estimation with progressive correlation modeling [2]. Both are relevant comparators.

Learned spatial transformation is established more broadly. Spatial Transformer Networks incorporate differentiable spatial transformations into learned networks [3]. PWC-Net combines pyramids, feature warping and cost volumes for optical flow [4]. These precedents motivate comparisons between global transformation, dense sampling and multiscale refinement. They do not imply that a task-trained sampling field is physically correct.

The intended distinction of SPAFNet is its combination of a semantic prior, four-scale progressive displacement estimation and gated bidirectional convolutional fusion. Whether the combination improves accuracy, efficiency or robustness requires controlled comparisons. This draft does not claim that semantic guidance or progressive alignment is itself new.

Zhai et al. also study semantic-guided deformable alignment and asymmetric fusion for this task [6]. A detailed implementation-level comparison is needed to distinguish SPAFNet’s contribution from this closely related work.

3 Method overview

3.1 Task and feature representation

Let I^r and I^t denote an RGB image and its thermal counterpart. The detector predicts a saliency map $\hat{Y}$ in a declared reference domain. An RGB reference is a convenient convention for this notation; the implementation and annotation protocol must confirm that convention.

Write the features at four scales as

$$\{R_i, T_i\}_{i=1}^4, \quad R_i, T_i \in \mathbb{R}^{H_i \times W_i \times C_i}, \quad (1)$$

where increasing i denotes a coarser scale. This common-grid notation is for exposition. It does not prescribe encoders, channel counts, parameter sharing, input resizing or the treatment of unequal image dimensions.

3.2 Shared semantic prior

A shared semantic prior provides task-level context to alignment and fusion. Conceptually, it can help associate an object region across modalities when local appearance differs. It supports relevant correspondence search without requiring similarity between RGB intensity and thermal intensity.

The supplied description does not identify the prior’s source features, construction layers, dimensionality or injection locations. Those choices matter: a coarse prior may suppress useful small-object evidence, while an incorrect semantic hypothesis may steer several scales toward the wrong region. Its effect therefore needs to be evaluated separately from the rest of the network.

3.3 Progressive dense feature sampling

At scale i , let $d_i(q)$ be the displacement at location q . The supplied design estimates it from inherited coarse flow, a local matching expectation and a learned residual. A compact dependency statement is

$$d_i = \mathcal{A}_i(\mathcal{U}_i(d_{i+1}), m_i, r_i; P), \quad i \in \{1, 2, 3\}, \quad (2)$$

where P is the semantic prior, $\mathcal{U}_i$ transfers a coarse field to the current scale, m_i denotes local matching information and r_i denotes a learned residual. Equation (2) records dependencies rather than asserting an unverified additive update rule. The coarsest field d_4 requires its own initialization; its construction, displacement units and the form of $\mathcal{A}_i$ remain unspecified.

Under a backward-sampling convention, thermal features queried by a reference location can be written as

$$\tilde{T}_i(q) = \mathcal{S}(T_i, q + d_i(q)), \quad (3)$$

with sampling operator $\mathcal{S}$. A complete implementation must define interpolation, coordinate normalization, flow rescaling, boundary handling and invalid samples. Without these definitions, apparently similar rules can have different spatial meanings.

Progression across scales provides a way to carry a broad displacement estimate into a finer search. Its capture range depends on feature resolution and matching support. Four scales alone do not guarantee recovery from arbitrary offsets, missing overlap or coarsest-scale errors.

3.4 Gated bidirectional fusion

The gate modulates the sampled thermal contribution before or within bidirectional convolutional fusion. Let g_i represent modulation weights compatible with $\tilde{T}_i$. The contribution can be summarized as

$$T_i^g = g_i \odot \tilde{T}_i, \quad (4)$$

where $\odot$ denotes elementwise multiplication with any required broadcasting. This shorthand leaves the gate’s range, spatial or channel structure, and parameterization to the implementation.

The supplied description identifies saliency supervision for the gate, without direct correspondence calibration. Accordingly, confidence-aware fusion describes the intended modulation function. It does not imply that a gate value is a calibrated probability of geometric correctness. The detector may suppress thermal evidence because it does not improve saliency, even when it is well aligned.

Bidirectional convolutional fusion is intended to permit reciprocal information exchange. The exact directions, branch updates, convolution kernels, fusion order and decoder connections are not recoverable from the abstract. In particular, this draft does not infer reverse flow estimation or cycle consistency from the word bidirectional.

3.5 Supervision

Saliency supervision links feature selection to the segmentation task. It can train a useful detector without ground-truth cross-modal flow, but does not uniquely determine that flow. Different fields

and gates may produce similar saliency masks.

The complete training objective has not been supplied. No backbone, optimizer, loss weights, initialization, image size, schedule, augmentation distribution or checkpoint-selection rule is asserted here. These details and the code version are necessary for reproduction.

4 Reported comparison and evaluation requirements

4.1 Available evidence

Table 1 records the two supplied UVT20K point estimates. PCNet’s 0.822 is identifiable in its original comparison table [2]; SPAFNet’s 0.887 has not been independently reproduced. Their numerical difference is 0.065. It is not, by itself, evidence of a controlled experimental improvement.

Table 1: Reported UVT20K maximum F_m values. Higher is better. Neither value was recomputed for this draft; common evaluation conditions have not been verified.

Method	Maximum $F_m \uparrow$	Evidence source
PCNet	0.822	Original paper and supplied text
SPAFNet	0.887	Supplied text only

Uncertainty, independent training-run counts and SPAFNet evaluation sample count are unavailable. No statistical ranking is implied.

The F_β family combines precision P_r and recall R_c :

$$F_\beta = \frac{(1 + \beta^2)P_r R_c}{\beta^2 P_r + R_c}. \quad (5)$$

A saliency evaluation precedent uses $\beta^2 = 0.3$ [5]. For the reported maximum F_m , the exact evaluator is still needed to establish thresholds, normalization, averaging order, handling of empty masks and choice of β . Maximizing an aggregated precision–recall curve need not equal averaging per-image maxima.

4.2 Dataset coverage

The supplied abstract states two naturally unaligned, three synthetically perturbed and three aligned datasets. It names only UVT20K. A relevant published protocol uses UVT20K and UVT2000, three un-VT variants and their three aligned counterparts [2]. This is context, not confirmation of the unnamed SPAFNet datasets. Their identities, versions, splits and results must be established from experiment records.

A complete evaluation should distinguish natural misalignment from synthetic perturbation. Known transformations support controlled tests of displacement behavior, whereas real scenes include occlusion and missing content that an image-plane perturbation may not reproduce. Aligned data can assess whether sampling and gating introduce a cost when spatial correspondence is already available.

4.3 Tests needed to support the claims

Controlled ablations should isolate the shared prior, inherited flow, matching expectation, residual update, gate and bidirectional fusion. Each comparison needs a common backbone, training data, evaluator and checkpoint-selection procedure. RGB-only, ungated fusion and no-resampling models would help identify the contributions of thermal evidence and alignment.

Geometric claims require geometric evidence, such as known-transform recovery or annotated cross-modal correspondences with visibility handled explicitly. Saliency scores and visually plausible feature maps alone cannot measure registration accuracy.

Reliability requires repeated runs and per-image predictions, with uncertainty analysis that accounts for related images where applicable. Robustness to new systems requires a declared held-out sensor or acquisition setup. Runtime and memory claims require hardware, resolution, batch size and timing conventions. None of these additional measurements is available in the supplied material.

5 Discussion and limitations

The proposed design has a coherent task-level motivation: sampling seeks relevant thermal evidence, while gating controls its influence on detection. Their interaction also creates ambiguity. A weak gate can hide a poor displacement estimate, and a strong semantic prior can encourage correct segmentation without establishing pointwise correspondence.

Modality-specific artifacts deserve separate evaluation. Thermal saturation, background heat, truncated objects and missing overlap can produce features with no useful RGB correspondence. Appropriate behavior may be to reduce their contribution rather than force an alignment. This is a hypothesis about intended behavior, not a reported finding.

The present evidence supports an architecture overview and a transparent record of one supplied comparison. It does not establish significance, causal module contributions, calibrated confidence, physically correct registration or generalization to unseen sensors. The complete tables, predictions, training records and implementation are needed for those conclusions.

Before public release, the authors should complete a current related-work review, verify the architecture against the implementation and confirm all result provenance. Code availability, funding, author contributions, conflicts of interest and rights to third-party material have not been supplied and are not inferred in this draft.

6 Conclusion

SPAFNet combines a shared semantic prior, progressive dense sampling at four scales and gated bidirectional convolutional fusion for unaligned RGB-T salient object detection. Its supplied UVT20K estimate is 0.887, compared with PCNet’s published 0.822. This draft retains the comparison with its evidential limits and identifies the details needed for a reproducible paper. Broader accuracy, registration and deployment claims remain to be established.

AI-assisted preparation

OpenAI Codex expanded the user-supplied title, author information and abstract into this draft, including its organization, schematic notation and editorial revisions. Codex did not conduct experiments or independently reproduce the reported SPAFNet measurements. The authors are responsible for the final manuscript and its claims.

References

- [1] K. Wang, D. Lin, C. Li, Z. Tu, and B. Luo. Alignment-Free RGBT Salient Object Detection: Semantics-guided Asymmetric Correlation Network and A Unified Benchmark. *arXiv:2406.00917*, 2024. arxiv.org/abs/2406.00917.
- [2] K. Wang, K. Chen, C. Li, Z. Tu, and B. Luo. Alignment-Free RGB-T Salient Object Detection: A Large-Scale Dataset and Progressive Correlation Network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(7):7780–7788, 2025. doi:10.1609/aaai.v39i7.32838. Preprint: arXiv:2412.14576.
- [3] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial Transformer Networks. *Advances in Neural Information Processing Systems*, 28, 2015. arXiv:1506.02025.
- [4] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8934–8943, 2018. arXiv:1709.02371.
- [5] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk. Frequency-tuned Salient Region Detection. *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. Author laboratory publication page.
- [6] S. Zhai, Y. Ba, and Y. Wang. Alignment-free RGB-T salient object detection: a semantic-guided deformable alignment and asymmetric fusion network. *Scientific Reports*, 2026. doi:10.1038/s41598-026-75471-7.