*electronics*

MDPI

*Article*

# Cascade Registration and Fusion for Unaligned Infrared and Visible Images in Autonomous Driving

**Long Xiao [1], Yidong Xie [2,*] and Chengda Yao [3]**

[1] Advanced Technology Research Institute, Beijing Institute of Technology, Jinan 250011, China; xiaolono1@163.com
[2] School of Mechanical-Electrical and Vehicle Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China
[3] Beijing Institute of Technology (Zhuhai), Beijing Institute of Technology, Zhuhai 519000, China; 3220245406@bit.edu.cn
* Correspondence: xieyd8112@163.com; Tel.: +86-13811025774

**Abstract**

Infrared and visible image fusion is a critical technology for enhancing the all-weather perception capabilities of autonomous driving systems. However, the inherent physical parallax of vehicle-mounted sensors combined with motion-induced vibrations makes it difficult to achieve strict alignment between the source images. Direct fusion of such misaligned pairs leads to ghosting artifacts, which significantly compromises driving safety. To address this challenge, this paper proposes a cascaded deep fusion framework tailored for autonomous driving scenarios. A dual-modal perception dataset is first constructed, incorporating realistic physical parallax and non-rigid deformations. Subsequently, a decoupled strategy is established, characterized by geometric correction followed by semantic fusion: the Static-Feature Recursive Registration (SFRR) network is utilized to explicitly correct the spatial misalignments caused by parallax, thereby establishing geometric consistency; then, the Hierarchical Invertible Block Fusion (HIBF) network achieves lossless integration of cross-modal features by combining spatial frequency separation with invertible interaction techniques. Experimental results demonstrate that the proposed method outperforms representative algorithms across several metrics, including Mutual Information (MI), Visual Information Fidelity (VIF), Structural Similarity (SSIM), and Correlation Coefficient (CC), producing high-quality fused images with clear structural definitions.

**Keywords:** autonomous driving; non-rigid registration; infrared and visible image fusion



Academic Editor: Beiwen Li

Received: 31 January 2026
Revised: 7 March 2026
Accepted: 13 March 2026
Published: 30 March 2026



## 1. Introduction

As autonomous driving technology advances toward high-order stages, environmental perception systems face significant challenges in handling extreme lighting, adverse weather, and complex dynamic road scenarios [1]. Single-modal visible light sensors primarily rely on ambient reflected light for imaging. Under conditions such as nighttime or tunnel entrances and exits, which are characterized by low illumination or drastic lighting fluctuations, the acquired texture information often undergoes severe degradation [2,3]. This results in a notable decline in the reliability of object detection and scene understanding [4]. In contrast, infrared cameras operate within the near-infrared spectrum, capturing reflected signals that are independent of visible illumination [5]. They possess an excellent ability to penetrate smoke and haze and are inherently less sensitive to sudden visible light fluctuations, enabling them to capture critical targets—such as pedestrians and

vehicles—around the clock [6]. Consequently, the complementarity of these two modalities is leveraged to effectively fuse salient infrared reflections with rich texture details [7]. This approach has emerged as a crucial technical trajectory for enhancing the all-weather robustness and environmental adaptability of autonomous driving perception systems, providing a more reliable data source for critical downstream tasks such as multi-modal object detection [8] and salient object detection [9].

Due to the separate positioning requirement of infrared and visible sensors separately to cover the required field of view, binocular parallax arising from the physical installation baseline is unavoidable. Furthermore, the rigid deformations triggered by road bumps and vehicle motion during travel make it difficult to align source image pairs strictly in the spatial domain [10–13].

Therefore, constructing a perception scheme capable of both accurately correcting geometric deviations and losslessly preserving key cross-modal information within dynamic, large-parallax vehicular environments remains an urgent and complex challenge.

To address the unique demands of non-rigid parallax and high-dynamic fusion in vehicular perception, this paper proposes a systematic solution. Our major contributions are summarized as follows:

1. We construct an unaligned dual-modal perception dataset tailored for autonomous driving. This dataset preserves the inherent physical parallax of sensors and the non-rigid displacements induced by road bumps, thereby filling the existing gap in benchmarks for unaligned image fusion within real-world dynamic scenarios. It provides reliable data support for research into robust perception mechanisms capable of resisting parallax interference.
2. A cascaded deep learning framework is proposed to decouple the perception task. By explicitly correcting spatial misalignments through a learned deformation field, the framework provides geometrically consistent inputs for subsequent processing stages, ensuring the physical fidelity and authenticity of the final fusion results.
3. Specialized registration and fusion networks are designed to handle structural features and hierarchical information. Specifically, the Static-Feature Recursive Registration (SFRR) network is introduced, which utilizes a multi-scale recursive mechanism to explicitly correct geometric deformations between source images, establishing pixel-level spatial consistency. Furthermore, the Hierarchical Invertible Block Fusion (HIBF) network is proposed; it leverages spatial frequency separation and invertible information interaction techniques to decouple high- and low-frequency information, achieving lossless complementarity between infrared spectral intensity and visible texture features.

## 2. Related Work

### *2.1. Deep Learning-Based Image Fusion*

With the rapid advancement of deep learning in computer vision, neural network-based image fusion methods have gradually replaced traditional multi-scale transform methods as the mainstream approach due to their powerful feature extraction and non-linear mapping capabilities [14–19]. Early research primarily relied on Convolutional Neural Network and Auto-Encoder architectures [20–23]. Specifically, DenseFuse introduced a fusion network based on dense connection blocks, which effectively preserves deep features of the source images by maximizing feature reuse within the encoder and reconstructing the image in the decoder [24]. RFN-Nest designed an end-to-end two-stage training strategy, utilizing a residual fusion network to replace handcrafted fusion rules while incorporating a Nest Connection architecture to leverage multi-scale deep features [25]. To enhance model generalization, U2Fusion proposed a unified unsuper-

vised framework capable of processing various fusion tasks, such as multi-modal and multi-exposure fusion, using a single model [26].

Despite the impressive performance of CNNs, they are limited by the local receptive fields of convolutional kernels, which results in deficiencies when capturing long-range dependencies. To address this issue, researchers have begun incorporating RNNs and attention mechanisms. Inspired by GRU, MemoryFusion designed a recurrent memory fusion unit and differential gating mechanisms to establish long-term dependencies across spatiotemporal dimensions, adaptively selecting high-value details and semantic information [27]. SwinFusion introduced the Swin Transformer into the fusion domain, leveraging the shifted window mechanism to achieve efficient multi-head self-attention and cross-attention; this enables effective global context modeling while retaining local details [28]. To address the issue of frequency domain information separation during feature extraction, MMIF-INet integrated traditional signal processing methods with deep networks by introducing the discrete wavelet transform. This method decomposes images into high-frequency details and low-frequency approximation components, constructing a multi-scale fusion module and corresponding hybrid loss functions. By employing refined processing in the frequency domain, it aims to preserve the structural details and spectral intensity of the source images, thereby avoiding information loss caused by downsampling [29]. NICE [30] utilizes an invertible framework with additive coupling layers to establish bijective mappings for lossless feature interaction in image fusion. RealNVP [31] achieves flexible and reversible information interaction for symmetric multi-modal feature integration through affine coupling layers and non-volume-preserving transformations.

However, most of the aforementioned methods operate under the assumption that the input infrared and visible images are strictly aligned at the pixel level. When confronted with parallax and displacements in real-world scenarios, directly applying these algorithms often results in severe ghosting artifacts and blurring in the fused outputs.

### 2.2. Cross-Modal Image Registration Challenges

Although existing deep fusion networks have made significant progress in feature extraction and reconstruction, their performance remains highly dependent on strict pixel-level alignment of the input images [32,33]. In practical applications such as autonomous driving, non-rigid spatial misalignments between source image pairs are inevitable due to the physical parallax between infrared and visible sensors, as well as dynamic vibrations during vehicle operation.

To obtain spatially consistent inputs, early studies frequently employed traditional registration techniques based on handcrafted features or mutual information as a preprocessing step [34]. However, constrained by the vast nonlinear radiometric differences between infrared and visible images, these methods struggle to extract stable consistent features [35].Furthermore, they are mostly predicated on rigid transformation assumptions, making them incapable of precisely modeling complex local non-rigid deformations caused by parallax [36,37]. This lack of preprocessing precision directly leads to residual spatial offsets, which in turn trigger severe ghosting artifacts in subsequent fusion results [38–40].

Given the limitations of traditional methods in handling cross-modal non-rigid deformations, utilizing data-driven deep features for spatial correction has become pivotal to addressing this issue. Representative architectures in this field include FlowNet [41], which utilizes an encoder–decoder structure with skip connections to directly regress dense displacement fields for end-to-end spatial alignment. PWC-Net [42] integrates feature pyramids, warping mechanisms, and cost volumes within a unified framework to achieve accurate coarse-to-fine spatial matching. Additionally, Deformable Convolutional Networks (DCN) [43] introduce learnable spatial offsets into convolutional sampling

grids to model adaptive local geometric deformations for precise alignment. Leveraging powerful nonlinear representation capabilities, deep networks can bridge the radiometric gap and capture the latent deep structural consistency between images. By constructing learnable alignment modules to explicitly model pixel-level correspondences, this strategy effectively overcomes geometric distortions caused by complex parallax and provides strictly aligned semantic features for the subsequent fusion network, thereby ensuring that the synthesized images possess both distinct geometric structures and high-quality information representation.

## 3. Proposed Method

In autonomous driving scenarios, the direct fusion of infrared and visible images affected by parallax leads to ghosting artifacts on road markings and blurred vehicle contours, which consequently interferes with the decision-making of perception algorithms. To address this issue, a cascaded deep learning framework is proposed in this paper, as illustrated in Figure 1.

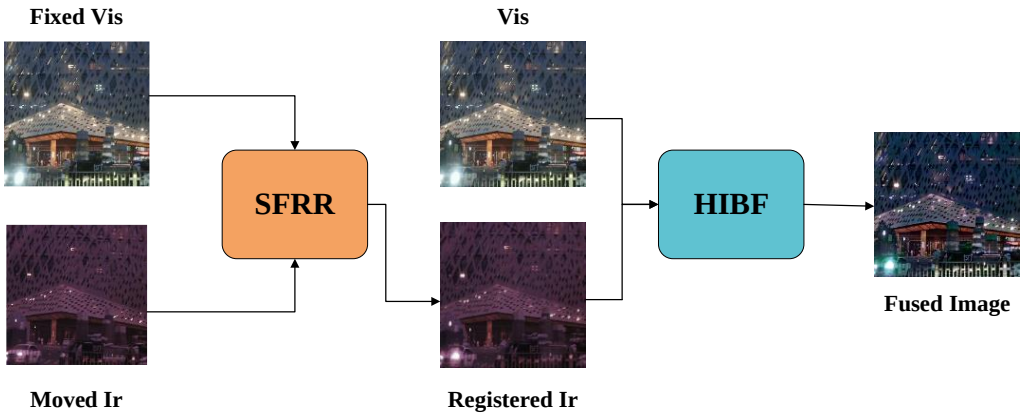


**Figure 1.** Overall Architecture of a Cascaded Deep Learning Framework.

The proposed framework explicitly decouples the fusion process of unaligned images into two independent stages: geometric correction and semantic integration. During the geometric correction stage, the SFRR network—mimicking the human visual alignment mechanism—explicitly computes pixel-level deformation fields. This process eliminates the non-rigid parallax triggered by sensor configurations and vehicle motion, thereby establishing precise spatial correspondences at the pixel level. Subsequently, in the semantic integration stage, the HIBF network facilitates the deep fusion of complementary information based on the achieved geometric consistency.

Let $I_{\text{ir}}$ denote the moving infrared image to be registered and $I_{\text{vis}}$ represent the fixed visible reference image. Initially, the registration module estimates a spatial transformation to align the infrared image with the visible reference:

$$I_{\text{ir}}^{\text{reg}} = \mathcal{F}_{SFRR}(I_{\text{ir}}, I_{\text{vis}}) \tag{1}$$

where $I_{\text{ir}}^{\text{reg}}$ denotes the registered infrared image. Subsequently, the fusion module integrates the complementary information from the aligned image pair to yields the final output $I_{\text{fused}}$:

$$I_{\text{fused}} = \mathcal{F}_{\text{HIBF}}\left(I_{\text{vis}}, I_{\text{ir}}^{\text{reg}}\right) \tag{2}$$

### *3.1. Static-Feature Recursive Registration*

To align the unaligned infrared image with the fixed visible reference image, we propose the Static-Feature Recursive Registration network. Rather than relying on volatile pixel intensities, the SFRR establishes correspondences based on stable structural context, thereby ensuring precise spatial alignment despite the substantial radiometric gap between the two modalities. As illustrated in Figure 2, it comprises two parallel feature extraction branches that map the input images into four-level pyramid feature spaces, thereby capturing geometric context across scales ranging from global to local. During the decoding stage, the network adopts a coarse-to-fine recursive strategy. Deformation Update Blocks (DUBs) are cascaded to stepwise regress the deformation field. The alignment process commences at the deepest and coarsest semantic level to recover large displacements, then recursively refines upward to the shallowest level to correct subtle structural deviations, ultimately yielding the registered infrared image.

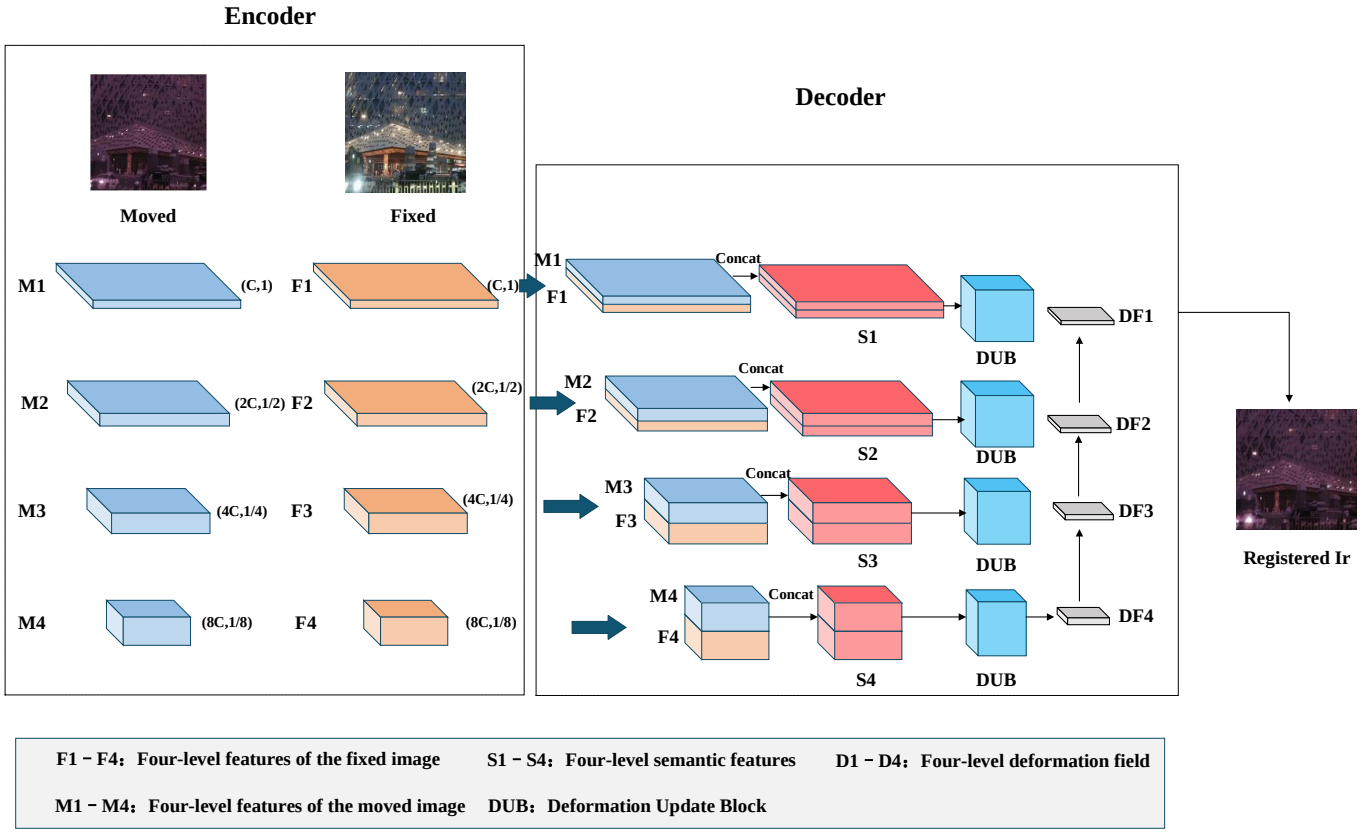


**Figure 2.** Overall Architecture of SFRR.

#### 3.1.1. Encoder

To capture semantic structures and local texture details, we employ a weight-sharing encoder. The encoder processes $I_{\text{vis}}$ and $I_{\text{ir}}$ independently across four hierarchical levels. Let $\text{Enc}^l(\cdot)$ denote the encoding operation at the $l$-th level. Consequently, two sets of feature pyramids are extracted:

$$F_l = \text{Enc}^l(I_{\text{vis}}), M_l = Enc^l(I_{\text{ir}}), l \in \{1,2,3,4\} \quad (3)$$

The spatial resolution is progressively reduced from the 1st to the 4th level, while the channel capacity increases accordingly. This multi-scale architecture allows for the simultaneous extraction of long-range geometric context and local textural intricacies. By prioritizing the extraction of shared structural representations that remain consistent across

distinct spectral domains, the encoder effectively filters out modality-specific radiometric fluctuations and sensor noise. This strategy ensures that the hierarchical features concentrate on underlying scene geometry rather than transient intensity variations. Consequently, it provides stable geometric anchors for the subsequent recursive registration stage.

#### 3.1.2. Decoder

The core of SFRR is the recursive decoding process, which utilizes DUBs to progressively refine the deformation field from the coarsest level $l$ to the finest level. At each hierarchical level $l$, the fixed and moving features $(F_l, M_l)$ are first concatenated to formulate a composite semantic feature volume $S_l$. The DUB is introduced as the pivotal regression unit; it is responsible for learning dense displacement information from the feature volume. More precisely, it constructs a regular sampling grid $G$ via meshgrid generation, onto which the predicted flow field is superposed to derive the target coordinates. Subsequently, a normalization operation is performed on these coordinates to map the displaced positions into the $[-1, 1]$ interval:

$$loc_{\text{norm}} = 2 \times \frac{loc_{\text{new}}}{size - 1} - 1 \tag{4}$$

where $x$ denotes the target pixel-level spatial coordinate derived by superposing the predicted displacement onto the sampling grid. *size* represents the scale of the feature map at the current hierarchical level: it refers to the width $W$ when calculating horizontal normalized coordinates and the height $H$ for vertical ones. $loc_{\text{norm}}$ is the resulting normalized coordinate that provides a standardized reference for the sampling process.

Subsequently, bilinear interpolation is employed to achieve the spatial alignment of features. As the coarsest level ($l = 4$) lacks prior knowledge, the initial deformation field is computed directly from the current-level features:

$$DF_4 = \text{DUB}_4(S_4) \tag{5}$$

For subsequent levels ($l < 4$), a residual learning strategy is adopted to capture finer-grained non-rigid deformations. The estimated flow field $DF_{l+1}$ from the preceding level is upsampled to match the current resolution, serving as the foundational guidance. Subsequently, the DUB focuses on the unaligned residuals at the current scale, estimating an incremental deformation to refine the alignment:

$$DF_l = \text{DUB}_l(S_l) + \text{Upsample}(DF_{l+1}) \tag{6}$$

Ultimately, by utilizing the finest flow field result $DF_1$ to warp the original input, the final registered infrared image $I_{\text{ir}}^{\text{reg}}$ is obtained.

### *3.2. Hierarchical Invertible Block Fusion*

Following geometric alignment, the Hierarchical Invertible Block Fusion network is employed to integrate multi-modal information. As illustrated in Figure 3, the framework utilizes a frequency-decoupling workflow to separately process texture details and structural intensities. The input source features are first processed by the Spatial Frequency Separator (SFS), which decomposes them into low-frequency (LF) components representing base energy and high-frequency (HF) components containing fine-grained edge details. To ensure the comprehensive preservation of critical foundational information, the LF components undergo deep interaction through Invertible Blocks. Finally, the calibrated LF features are recombined with the preserved HF features, and a Convolutional Cascade module

reconstructs these high-dimensional features back into the image domain to generate the high-quality final fused output.

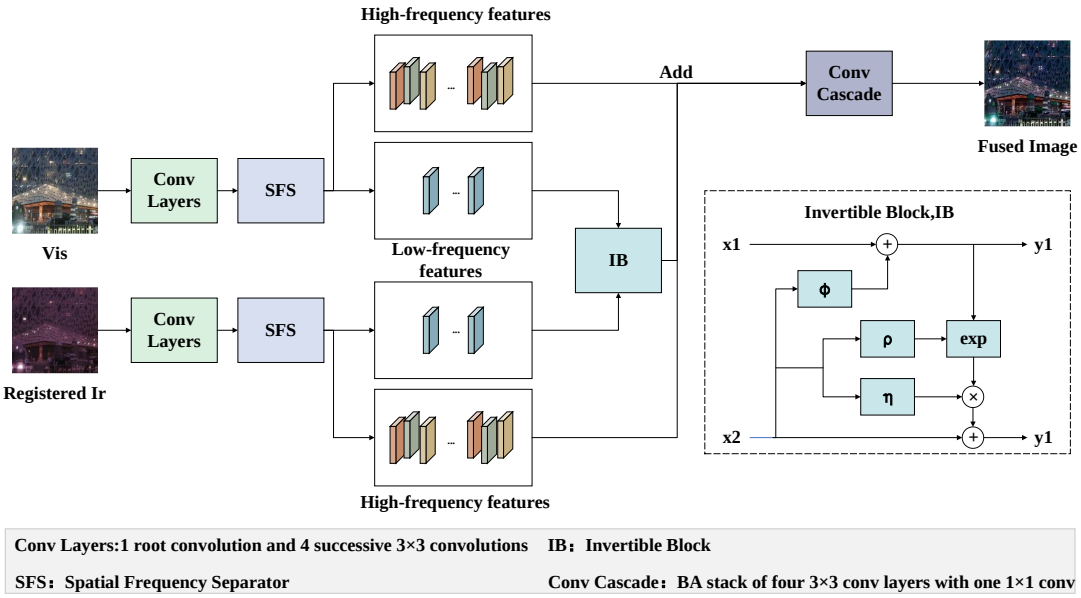


**Figure 3.** Overall Architecture of HIBF.

### 3.2.1. Spatial Frequency Separation

As illustrated in Figure 4, the SFS module is introduced to explicitly decouple geometric structures from texture details. Let $X$ denote the input feature map. A Gaussian blur kernel is applied to extract the fundamental coarse information $X_{base}$, while the HF components are derived via residual subtraction. Subsequently, both components undergo downsampling to construct a pyramid representation.

$$X_{\downarrow}^{LF} = \text{AvgPool}(X_{base}), X_{\downarrow}^{HF} = \text{AvgPool}\left(X_{hf}\right) \tag{7}$$

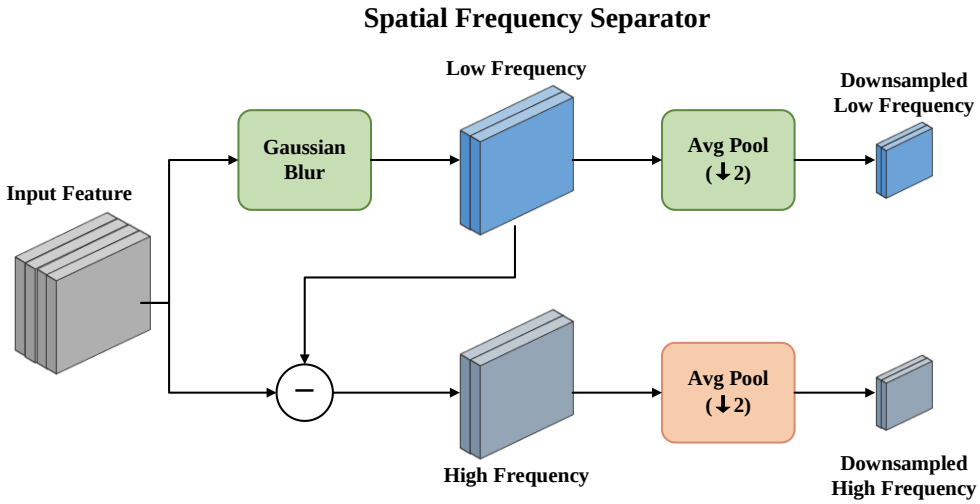


**Figure 4.** Overall Architecture of SFS.

Here, y represent the LF base information and HF texture details derived via Gaussian blurring and residual operations, respectively. AvgPool$(\cdot)$ denotes the average pooling operation, utilized to regulate the spatial resolution of the feature maps. $X_{\downarrow}^{LF}$ and $X_{\downarrow}^{HF}$ are the resulting downsampled LF and HF components. Through this mechanism, the network

is capable of processing infrared reflection and structural information in parallel paths, thereby mitigating mutual interference between distinct feature types.

#### 3.2.2. Invertible Information Interaction

To facilitate deep semantic integration between the decoupled low-frequency components, we employ Invertible Blocks based on an affine coupling mechanism. Given the input visible features $x_1$ and infrared features $x_2$, the information interaction is formulated through a dual-stream transformation as follows:

$$y_1 = x_1 + \phi(x_2) \tag{8}$$

$$y_2 = x_2 \odot \exp(\rho(y_1)) + \eta(y_1) \tag{9}$$

where $\phi$, $\rho$, and $\eta$ denote arbitrary learnable sub-networks that extract cross-modal context. In this configuration, the visible components $x_1$ are first modulated by the infrared information through an additive operation in Equation (8). Subsequently, the infrared feature $x_2$ undergoes a scaling transformation via the element-wise multiplication $\odot$ and a translational shift based on the updated visible state $y_1$ in Equation (9). This affine coupling structure ensures that the multi-modal features are thoroughly fused while maintaining strict numerical invertibility, which prevents the loss of salient information during the deep interaction process.

#### 3.2.3. Feature Reconstruction and Cascade

Upon completing deep feature interaction within the frequency domain, the final step involves reconstructing a fused image that conforms to human visual perception. This stage encompasses frequency recombination and spatial mapping.

Initially, the refined features from the IB modules are recombined with the preserved HF details. Since the HF components, $HF_{\text{vis}}$ and $HF_{\text{ir}}$, carry complementary edge information, we utilize element-wise addition to merge them, thereby ensuring the preservation of distinct structural details. The aggregate reconstructed feature map, $F_{\text{rec}}$, is derived as follows:

$$F_{\text{rec}} = \text{IB}_{\text{out}}(LF_{\text{vis}}, LF_{\text{ir}}) + (HF_{\text{vis}} + HF_{\text{ir}}) \tag{10}$$

Subsequently, $F_{\text{rec}}$ is fed into the convolutional cascade module. This module comprises a series of $3 \times 3$ convolutional filters and $1 \times 1$ mapping layers, designed to decode high-dimensional integrated features back into the image domain. By progressively refining pixel intensities, the convolutional cascade synthesizes the final fused image $I_{\text{fused}}$. The resulting image effectively harmonizes the high infrared intensity contrast from the infrared source with the rich textural clarity from the visible source, while minimizing blurring artifacts.

### *3.3. Optimization Strategy and Loss Functions*

The proposed framework encompasses two distinct objectives: geometric alignment and semantic integration. To fulfill these requirements, we adopt a progressive, step-wise optimization strategy.

#### 3.3.1. Stage I: Registration Optimization

In the initial stage, the SFRR network is optimized to estimate precise deformation fields. Denoting $\theta_{\text{reg}}$ as the set of learnable parameters for the registration network, the primary objective is to minimize the aggregate registration loss $\mathcal{L}_{\text{reg}}$, formulated as:

$$\min_{\theta_{\text{reg}}} \mathcal{L}_{\text{reg}} = \mathcal{L}_{\text{sim}}\left(I_{\text{vis}}, I_{\text{ir}}^{\text{reg}}\right) + \lambda_{\text{smooth}} \mathcal{L}_{\text{smooth}}(\phi) \tag{11}$$

where $I_{\text{vis}}$ represents the fixed visible image, and $I_{\text{ir}}^{\text{reg}}$ denotes the registered infrared image. $\phi$ signifies the predicted deformation field, while $\lambda_{\text{smooth}}$ is the hyperparameter utilized to balance the trade-off between the similarity term and the regularization term.

To account for nonlinear intensity variations between infrared and visible modalities, we utilize the Local Normalized Cross-Correlation (LNCC) as the similarity measure. For a given pixel $p$, the LNCC is computed over a local neighborhood window $\Omega_p$ to ensure structural consistency across different sensors. The loss function is formally defined as:

$$\mathcal{L}_{\text{sim}}(I,J) = 1 - \frac{1}{N}\sum_{p} \frac{\sum_{p_i\in\Omega_p}\left(I(\,p_i\right) - \overline{I}(p))\left(J(p_i) - \overline{J}(p)\right)}{\sqrt{\sum_{p_i\in\Omega_p}\left(I(p_i) - \overline{I}(p)\right)^2}\sqrt{\sum_{p_i\in\Omega_p}\left(J(p_i) - \overline{J}(p)\right)^2}} \tag{12}$$

where $I$ and $J$ represent the fixed visible image and the registered infrared image, respectively. $N$ signifies the total number of pixels in the image. $\Omega_p$ denotes the local neighborhood window centered at pixel $p$. $\overline{I}(p)$ and $\overline{J}(p)$ refer to the local mean intensities of the respective images within the window $\Omega_p$. By maximizing the squared correlation, the LNCC allows the network to focus on structural patterns while remaining invariant to absolute pixel intensity fluctuations, ensuring robust spatial consistency across modalities.

To promote physically plausible deformations and prevent topological folding, we apply diffusion regularization to the spatial gradient of the deformation field $\phi$:

$$\mathcal{L}_{\text{smooth}}(\phi) = \sum_{p} \| \nabla\phi(p) \|_2^2 = \sum_{p}\left(\left(\frac{\partial\phi(p)}{\partial x}\right)^2 + \left(\frac{\partial\phi(p)}{\partial y}\right)^2\right) \tag{13}$$

where $\nabla$ denotes the vector gradient operator, while $\frac{\partial\phi(p)}{\partial x}$ and $\frac{\partial\phi(p)}{\partial y}$ represent the partial derivatives of the flow field at position $p$ along the horizontal and vertical directions, respectively.

#### 3.3.2. Stage II: Fusion Optimization

In the second stage, the parameters $\theta_{reg}$ are frozen. The pre-aligned infrared image $I_{ir}^{reg}$ and the visible image $I_{vis}$ are utilized as inputs to optimize the parameters $\theta_{fus}$ of the HIBF network.

$$\min_{\theta_{\text{fus}}} \mathcal{L}_{\text{fus}} = \alpha\mathcal{L}_{\text{int}} + \beta\mathcal{L}_{\text{ssim}} + \gamma\mathcal{L}_{\text{grad}} \tag{14}$$

Here, The expression $\min_{\theta_{\text{fus}}}\mathcal{L}_{\text{fus}}$ denotes the optimization objective, $\alpha$, $\beta$, and $\gamma$ denote the trade-off weights that regulate the relative contributions of the intensity loss, SSIM loss, and gradient loss, respectively.

To ensure that the fused image $I_{\text{fused}}$ preserves the most critical infrared reflection and reflected light information, we constrain the fused image to approximate the element-wise maximum of the source inputs $\mathcal{L}_{\text{int}}$:

$$\mathcal{L}_{\text{int}} = \frac{1}{HW} \| I_{\text{fused}} - M \|_1 \tag{15}$$

where $H$ and $W$ denote the height and width of the image, $\| \cdot \|_1$ represents the L1-norm, and $M = \max\left(I_{\text{ir}}^{\text{reg}}, I_{\text{vis}}\right)$ signifies the pixel-wise maximum intensity map derived from the source images.

To maintain perceptual integrity, the Structural Similarity Index $\mathcal{L}_{ssim}$ is utilized. This loss ensures that the fused image remains structurally consistent with both source images:

$$\mathcal{L}_{\text{ssim}} = \frac{1}{2}(1 - \text{SSIM}(I_{\text{fused}}, I_{\text{vis}})) + \frac{1}{2}\left(1 - \text{SSIM}\left(I_{\text{fused}}, I_{\text{ir}}^{\text{reg}}\right)\right) \tag{16}$$

To compel the network to retain fine-grained textural details and sharp edges, we minimize the discrepancy between the gradient map of the fused image and the joint maximum gradient of the source images:

$$\mathcal{L}_{\text{grad}} = \frac{1}{HW} \left\| \left| \nabla I_{\text{fused}} \right| - \max\left( \left| \nabla I_{\text{ir}}^{\text{reg}} \right|, \left| \nabla I_{\text{vis}} \right| \right) \right\|_1 \tag{17}$$

Here, $\nabla$ denotes the Sobel gradient operator utilized to extract high-frequency edge information, while $|\cdot|$ represents the absolute value operation.

The gradient loss in Equation (17) specifically incentivizes the preservation of sharp edges and fine-grained textures in the fused output. By minimizing the pixel-wise discrepancy between the gradient of the fused image and the joint maximum gradient derived from the source pair, the network is constrained to inherit the most salient structural transitions from either the infrared or visible modality. In autonomous driving scenarios, this mechanism is crucial for retaining the distinct boundaries of critical objects, such as lane markings and pedestrian silhouettes, which effectively prevents the blurring artifacts often caused by cross-modal feature integration.

## 4. Experiment

To validate the efficacy of the proposed cascaded framework in non-rigidly misaligned infrared and visible image fusion tasks, we designed and conducted extensive comparative experiments. This section initially elaborates on the collection and construction process of the self-built dataset, along with specific implementation details of the training process. Subsequently, four key metrics employed for the objective evaluation of fusion quality are introduced. Finally, the proposed method is subjected to a comprehensive quantitative and qualitative comparative analysis against current mainstream existing algorithms.

### *4.1. Dataset and Implementation Details*

To construct a high-quality benchmark that closely reflects real-world application scenarios, a dedicated dataset was produced by establishing a dual-band image synchronous acquisition system. The core hardware of this system consists of two KM2010-3660 industrial-grade camera modules with identical specifications, both of which are equipped with 1/2.7-inch OV2710 CMOS sensors. The visible light camera is responsible for capturing ambient color and textural details, while the infrared camera operates within the 700–1100 nm spectral band to capture infrared reflection characteristics. Both cameras feature a pixel size of 3.0 μm and are rigidly fixed to ensure the stability of the physical baseline, with both configured to output a resolution of 1920 × 1080.

The self-built dataset provides dual-modal image sequences that incorporate realistic physical parallax and non-rigid deformations caused by vehicle dynamics. The horizontal parallax within the dataset ranges from approximately 2 pixels for distant backgrounds beyond 100 m to 45 pixels for targets as close as 5 m. Furthermore, the dataset covers stochastic vibration jitter with measured displacement amplitudes ranging from 5 to 18 pixels across the modalities. To ensure the reliability of the spatial correspondences, the alignment of the dataset was verified through a manual protocol involving the annotation of prominent landmarks on a representative subset of 100 image pairs. The quantitative assessment yielded a Mean Registration Error of 1.26 pixels, confirming that the spatial consistency of the dataset is sufficient to support high-quality fusion results without visible ghosting artifacts.

To simulate actual operating environments for autonomous driving, the data acquisition scenes encompass complex road conditions, moving vehicles, and dynamic lighting variations. We extracted 1954 pairs of images from the captured video streams. Following a ratio of 7:1.5:1.5, the dataset was randomly partitioned into three subsets: 1368 pairs serve

as the training set for network parameter optimization, 293 pairs constitute the validation set for monitoring model convergence and hyperparameter tuning, and the remaining 293 pairs are utilized as the test set for final performance evaluation.

All code for the experiments was implemented based on the PyTorch 2.2.2 deep learning framework. Regarding the training strategy, a phased optimization approach was adopted: the training set was first utilized for the unsupervised training of the SFRR network until loss convergence, after which the parameters of the registration module were frozen to facilitate end-to-end weight updates for the HIBF network. Prior to being fed into the networks, all input images were converted to grayscale and normalized to the range of 0 to 1. The experiments were conducted on a workstation equipped with high-performance NVIDIA GPUs. Parameter updates were performed using the Adam optimizer, with the initial learning rate set to $1 \times 10^{-4}$. An adaptive decay strategy was implemented based on variations in validation loss to ensure that the model achieved optimal convergence while maintaining robust performance.

### *4.2. Evaluation Metrics*

To comprehensively and objectively evaluate the quality of the fused images, four quantitative metrics were selected for this study [44–47]. MI is based on the concept of entropy in information theory and measures the total amount of information transferred from the source inputs to the fusion result by calculating the statistical dependence between the fused and source images. VIF leverages natural scene statistical models and the perceptual characteristics of the human visual system to quantify the information fidelity of the fused image relative to the source images. SSIM decomposes the measurement of image similarity into three independent comparison components—luminance, contrast, and structure—to assess the degree to which the fusion result preserves the structural features of the source scene. CC evaluates the statistical association between the fused and source images in terms of pixel intensity distribution and spectral characteristics by calculating their linear correlation.

### *4.3. Comparison with Existing Competitive Methods*

To demonstrate the advanced performance of the proposed method, we compared our framework with six representative methods, encompassing both traditional fusion algorithms and deep learning-based schemes. Table 1 presents the average objective metrics for all methods evaluated on the test set. Quantitative results indicate that the proposed method achieves highly competitive performance across all four metrics. Specifically, it reached a value of 6.321 for MI, an improvement of approximately 13.8 percent relative to the second-best method, SwinFusion. In terms of VIF, the proposed method yielded a score of 1.305, outperforming all other comparative algorithms. Furthermore, an SSIM of 0.910 and a CC of 0.958 demonstrate that our approach possesses significant advantages in preserving the geometric structures of road scenes, lane marking edges, and vehicle outlines.

To qualitatively evaluate the fusion performance under unaligned conditions, we provide a detailed visual comparison in Figure 5, with red-box annotations and corresponding enlarged patches focusing on a vehicle. Due to the inherent physical parallax between sensors, there is a spatial discrepancy between the infrared and visible sources. As observed in the zoomed-in patches, representative methods such as SwinFusion, RFN-Nest, and MMIF exhibit severe ghosting artifacts and blurred contours around the vehicle's rear and license plate. This is because these methods directly fuse features without spatial correction, leading to the superposition of misaligned targets. In contrast, the fused image generated by our proposed method displays a clean structure with sharp boundaries. By

explicitly estimating the deformation field through the SFRR module, our framework successfully eliminates ghosting, producing the most perceptually natural and geometrically accurate results.

**Table 1.** Quantitative comparison results of the proposed framework and six representative methods evaluated on the test set.

| Method | MI | VIF | SSIM | CC |
|---|---|---|---|---|
| MMIF | 4.063 | 1.151 | 0.794 | 0.920 |
| DenseFuse | 5.055 | 0.661 | 0.837 | 0.948 |
| MemoryFusion | 4.197 | 0.588 | 0.756 | 0.909 |
| U2Fusion | 2.69 | 0.072 | 0.455 | 0.651 |
| SwinFusion | 5.553 | 0.827 | 0.802 | 0.934 |
| RFNNest | 4.872 | 0.608 | 0.758 | 0.942 |
| Ours | 6.321 | 1.305 | 0.910 | 0.958 |

### *4.4. Ablation Study*

To verify the efficacy of the core modules within the proposed cascaded framework and their individual contributions to the final fusion performance, a series of ablation experiments were conducted on the test set. We constructed four distinct variant models to evaluate the roles of SFRR and HIBF, as well as the internal components of the fusion network. These variants include the "w/o SFRR" model, which removes the registration module to process unaligned source images; the "SFRR-Single + HIBF" model, which replaces the recursive DUB mechanism with a single-step registration at the highest resolution to verify the value of the recursive architecture; the "w/o HIBF" model, which aligns source images using SFRR but applies a baseline average fusion strategy; the "w/o SFS" model, which excludes the frequency separation mechanism; and the "w/o IB" model, where Invertible Blocks are replaced by standard residual convolutional blocks. All variants were retrained and tested under identical hyperparameters, with results summarized in Table 2.

**Table 2.** Ablation study of different modules in the proposed framework.

| Method | MI | VIF | SSIM | CC |
|---|---|---|---|---|
| w/o SFRR | 5.742 | 1.035 | 0.812 | 0.884 |
| SFRR-Single + HIBF | 5.903 | 1.142 | 0.824 | 0.901 |
| w/o HIBF | 5.211 | 0.894 | 0.889 | 0.935 |
| w/o SFS | 6.085 | 1.214 | 0.887 | 0.941 |
| w/o IB | 5.953 | 1.182 | 0.895 | 0.938 |
| Ours | 6.321 | 1.305 | 0.910 | 0.958 |

The experimental results indicate that the complete computational framework achieves the highest performance across all metrics. First, a comparison between "w/o SFRR" and the full model reveals that a lack of geometric alignment leads to a degradation in performance, particularly in the SSIM and CC metrics, which decreased by approximately 44% and 29%, respectively. This indicates that without addressing spatial misalignment, ghosting artifacts occur, making SFRR the foundation for fusion quality. Second, although the "SFRR-Single" model improves the results compared to the unaligned version, it still underperforms compared to our recursive model in SSIM and CC. This demonstrates that the multi-scale recursive mechanism effectively handles large-parallax and non-rigid deformations that a single-step registration fails to correct. Furthermore, the "w/o HIBF" experiment shows that after alignment, a baseline fusion strategy leads to lower MI and VIF scores compared to the proposed HIBF network. Furthermore, the performance declines in the "w/o SFS" and "w/o IB" models confirm that frequency decoupling and invertible

interaction are the mechanisms through which HIBF maintains information fidelity. In summary, SFRR ensures spatial geometric accuracy, while HIBF enables the integration of semantic features.

To evaluate the effectiveness of the proposed SFRR, we conducted a component substitution experiment. While keeping the HIBF module constant, the SFRR was replaced by three representative alignment architectures: FlowNetS, PWC-Net, and a DCN-based baseline. The quantitative results on the test set are summarized in Table 3.

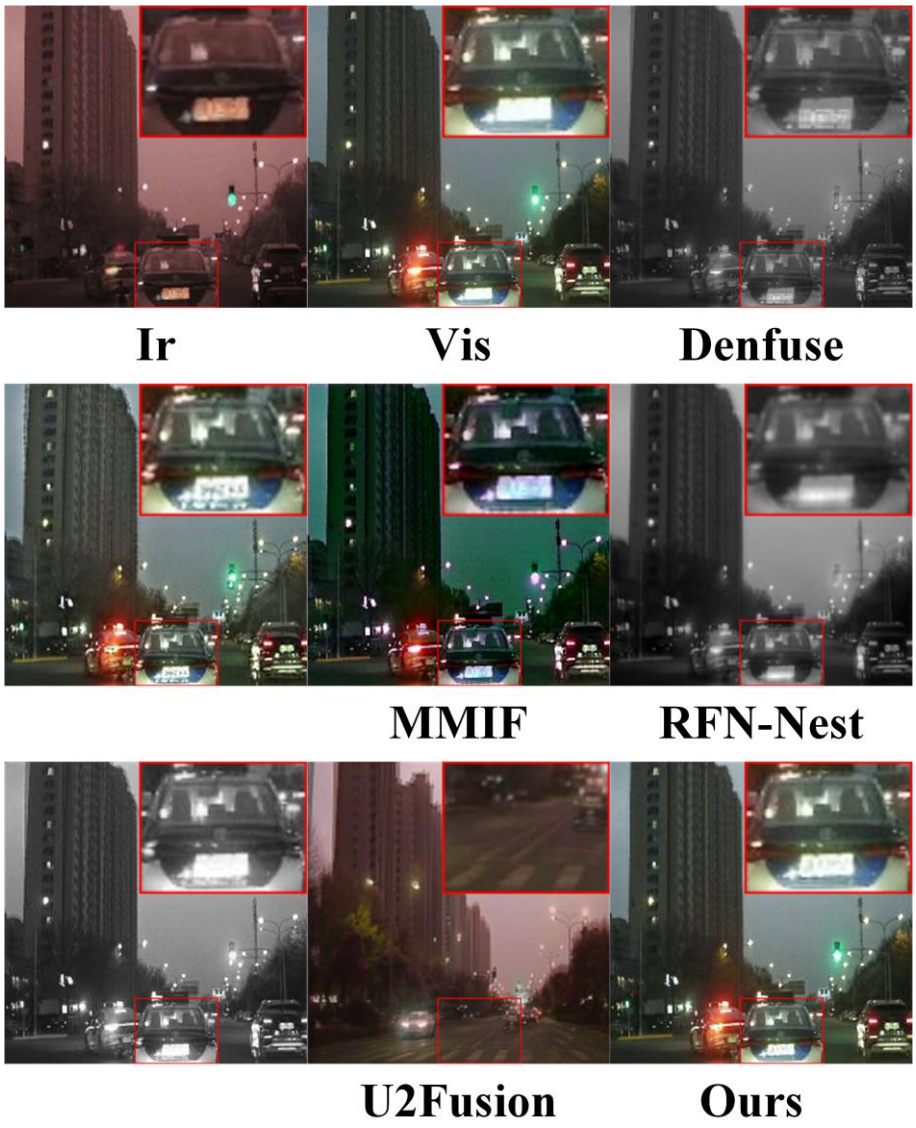


**Figure 5.** Qualitative comparison of the proposed method against several representative algorithms on an unaligned image pair from the autonomous driving dataset. Red frames highlight the regions where our method achieves better performance.

**Table 3.** Performance comparison of different registration modules (Fusion module fixed as HIBF).

| Method | MI | VIF | SSIM | CC |
|---|---|---|---|---|
| FlowNetS + HIBF | 4.542 | 0.785 | 0.762 | 0.884 |
| PWC-Net + HIBF | 4.924 | 0.908 | 0.814 | 0.912 |
| DCN-Baseline + HIBF | 5.235 | 0.942 | 0.835 | 0.918 |
| Ours | 6.321 | 1.305 | 0.910 | 0.958 |

The results demonstrate that our full model (SFRR+HIBF) achieves the highest scores. When PWC-Net is employed, the SSIM drops to 0.814, indicating that its intensity-based cost volume struggles to handle the non-linear radiometric gap between infrared and visible light, leading to residual ghosting artifacts. Similarly, the DCN baseline yields inferior metrics because its local offset learning lacks the global-to-local recursive refinement necessary for large cross-modal parallax. The superiority of SFRR stems from its weight-sharing recursive strategy, which focuses on semantic structural consistency rather than pixel-intensity correlation, effectively bridging the spectral gap.

Furthermore, we analyzed the innovation of the HIBF module by comparing it with classical invertible neural network (INN) structures. In this experiment, the registration module was fixed as SFRR, while the HIBF was replaced by NICE and RealNVP coupling layers, which operate on the full spectrum without frequency decoupling. As shown in Table 4, HIBF significantly outperforms the classical INN methods, particularly in MI and SSIM.

**Table 4.** Performance comparison of different invertible fusion strategies (Registration module fixed as SFRR).

| Method | MI | VIF | SSIM | CC |
|---|---|---|---|---|
| SFRR + NICE | 4.652 | 0.824 | 0.782 | 0.814 |
| SFRR + RealNVP | 5.315 | 1.025 | 0.835 | 0.886 |
| Ours | 6.321 | 1.305 | 0.910 | 0.958 |

While NICE and RealNVP maintain mathematical invertibility, their global interaction mechanism fails to isolate the low-frequency spectral energy of infrared images. Consequently, the rich high-frequency textures of visible light are often suppressed during the affine coupling process, leading to lower VIF and SSIM scores. In contrast, HIBF introduces hierarchical frequency separation, applying invertible interactions exclusively to the low-frequency components while preserving textures via a high-frequency bypass. This ensures that the fused images inherit high-contrast infrared intensities without losing the intricate environmental details.

### *4.5. Computational Efficiency Analysis*

In autonomous driving scenarios, inference latency and computational efficiency are as critical as perception accuracy. To systematically evaluate the real-time processing capabilities of our proposed method, we conducted a computational cost analysis. Table 5 presents the total number of learnable parameters (Params), floating-point operations (FLOPs), average inference time per image pair, and frames per second (FPS) for various methods. All algorithms were evaluated on images resized to 640 $\times$ 480 using a single NVIDIA RTX 3090 GPU.

**Table 5.** Computational efficiency comparison of different methods on an NVIDIA RTX 3090 GPU.

| Method | Parameters (M) | FLOPs (G) | Inference Time (s) | FPS |
|---|---|---|---|---|
| DenseFuse | 0.07 | 13.2 | 0.015 | 66.6 |
| U2Fusion | 0.66 | 42.5 | 0.028 | 35.7 |
| RFN-Nest | 6.24 | 115.8 | 0.072 | 13.8 |
| MMIF | 14.50 | 88.3 | 0.051 | 19.6 |
| SwinFusion | 2.85 | 165.4 | 0.124 | 8.0 |
| MemoryFusion | 8.12 | 102.6 | 0.065 | 15.3 |
| Ours | 5.98 | 78.4 | 0.031 | 32.2 |

As demonstrated in Table 5, while lightweight models straightforwardly fusing unaligned images (e.g., DenseFuse and U2Fusion) achieve higher FPS, they suffer from severe ghosting artifacts when facing complex non-rigid parallax. Conversely, Transformer-based methods like SwinFusion inevitably incur heavy computational burdens, yielding only 8.0 FPS, making them unsuitable for real-time vehicular systems. By rationally decoupling the task into geometric correction and semantic integration, our framework limits the required parameters to 5.98 M. Although the cascaded architecture introduces slightly more latency than standalone lightweight networks, our method still achieves an impressive inference speed of 0.031 s per frame. This effectively satisfies the real-time requirement for continuous autonomous driving perception, securing a trade-off between geometric stability, information fidelity, and execution efficiency.

## 5. Conclusions

In this paper, a cascaded deep learning framework is proposed to address the non-rigid misalignment issue caused by physical parallax between infrared and visible images in vehicle perception systems, utilizing a self-built dataset that simulates autonomous driving road scenarios. Initially, the source images are processed by a structural feature registration network, which calculates the deformation field through a multi-scale recursive mechanism to achieve spatial correction. Subsequently, the geometrically consistent image pairs are fed into the fusion network, where they undergo spatial frequency separation, invertible information interaction, and feature cascaded reconstruction modules sequentially to generate the final fused image. Experimental results demonstrate that the proposed method achieves superior performance in terms of MI, VIF, SSIM, and CC, with favorable visual outcomes, thereby validating the efficacy of the cascaded architecture in handling non-rigidly misaligned autonomous driving data.

Future work will focus on the lightweight optimization of the network structure to reduce computational load and enhance inference speed, aiming to meet the real-time processing demands of practical vehicle-mounted embedded systems. Furthermore, we intend to explore the application value of the fused images in downstream all-weather autonomous driving object detection tasks.

**Author Contributions:** Conceptualization, L.X. and Y.X.; methodology, L.X.; software, L.X.; validation, L.X. and Y.X.; formal analysis, L.X.; investigation, L.X.; resources, C.Y.; data curation, C.Y.; writing—original draft preparation, L.X.; writing—review and editing, L.X. and Y.X.; visualization, L.X.; supervision, Y.X.; project administration, Y.X.; funding acquisition, Y.X. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Research on Intelligent driving vehicle decision-making based on multi-source data fusion and passenger ride experience, grant number ZR2023MF0766. This project is supported by the Advanced Technology Research Institute, Beijing Institute of Technology.

**Data Availability Statement:** The original contributions presented in the study are included in the article; further inquiries can be directed to the corresponding author.

**Conflicts of Interest:** The authors declare that they have no known competing financial interests or personal relationships.

## References

1. Yurtsever, E.; Lambert, J.; Carballo, A.; Takeda, K. A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access* **2020**, *8*, 58443–58469. [CrossRef]
2. Gao, C.; Wang, G.; Shi, W.; Wang, Z.; Chen, Y. Autonomous driving security: State of the art and challenges. *IEEE Internet Things J.* **2021**, *9*, 7572–7595. [CrossRef]

3. Zhao, J.; Zhao, W.; Deng, B.; Wang, Z.; Zhang, F.; Zheng, W.; Cao, W.; Nan, J.; Lian, Y.; Burke, A.F. Autonomous driving system: A comprehensive survey. *Expert Syst. Appl.* **2024**, *242*, 122836. [CrossRef]
4. Bachute, M.R.; Subhedar, J.M. Autonomous driving architectures: Insights of machine learning and deep learning algorithms. *Mach. Learn. Appl.* **2021**, *6*, 100164. [CrossRef]
5. Zhou, W.; Lin, X.; Lei, J.; Yu, L.; Hwang, J.-N. MFFENet: Multiscale feature fusion and enhancement network for RGB–thermal urban road scene parsing. *IEEE Trans. Multimed.* **2021**, *24*, 2526–2538. [CrossRef]
6. Dolatyabi, P.; Regan, J.; Khodayar, M. Deep Learning for Traffic Scene Understanding: A Review. *IEEE Access* **2025**, *13*, 13187–13237. [CrossRef]
7. Monninger, T.; Schmidt, J.; Rupprecht, J.; Raba, D.; Jordan, J.; Frank, D.; Staab, S.; Dietmayer, K. Scene: Reasoning about traffic scenes using heterogeneous graph neural networks. *IEEE Robot. Autom. Lett.* **2023**, *8*, 1531–1538. [CrossRef]
8. Liu, J.; Fan, X.; Huang, Z.; Wu, G.; Liu, R.; Zhong, W.; Luo, Z. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5802–5811.
9. Tang, H.; Li, Z.; Zhang, D.; He, S.; Tang, J. Divide-and-conquer: Confluent triple-flow network for RGB-T salient object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *47*, 1958–1974. [CrossRef]
10. Zhou, X.; Lin, Z.; Shan, X.; Wang, Y.; Sun, D.; Yang, M.-H. Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 21634–21643.
11. Fang, J.; Wang, F.; Xue, J.; Chua, T.-S. Behavioral intention prediction in driving scenes: A survey. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 8334–8355. [CrossRef]
12. Awan, M.; Whangbo, T.K.; Shin, J. Deep learning Methods for Autonomous Driving Scene Understanding Tasks: A Review. *Expert Syst. Appl.* **2025**, *287*, 128098. [CrossRef]
13. Rao, Y.R.; Prathapani, N.; Nagabhooshanam, E. Application of normalized cross correlation to image registration. *Int. J. Res. Eng. Technol.* **2014**, *3*, 12–16. [CrossRef]
14. Tang, L.; Deng, Y.; Ma, Y.; Huang, J.; Ma, J. SuperFusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA J. Autom. Sin.* **2022**, *9*, 2121–2137. [CrossRef]
15. Li, S.; Kang, X.; Fang, L.; Hu, J.; Yin, H. Pixel-level image fusion: A survey of the state of the art. *Inf. Fusion* **2017**, *33*, 100–112. [CrossRef]
16. Zhao, F.; Zhang, C.; Geng, B. Deep multimodal data fusion. *ACM Comput. Surv.* **2024**, *56*, 216. [CrossRef]
17. Liu, X.; Liu, Q.; Wang, Y. Remote sensing image fusion based on two-stream fusion network. *Inf. Fusion* **2020**, *55*, 1–15. [CrossRef]
18. Karim, S.; Tong, G.; Li, J.; Qadir, A.; Farooq, U.; Yu, Y. Current advances and future perspectives of image fusion: A comprehensive review. *Inf. Fusion* **2023**, *90*, 185–217. [CrossRef]
19. Tang, L.; Xiang, X.; Zhang, H.; Gong, M.; Ma, J. DIVFusion: Darkness-free infrared and visible image fusion. *Inf. Fusion* **2023**, *91*, 477–493. [CrossRef]
20. Blum, R.S.; Xue, Z.; Zhang, Z. An Overview of lmage Fusion. In *Multi-Sensor Image Fusion and Its Applications*; CRC Press: Boca Raton, FL, USA, 2018; pp. 1–36.
21. Zhang, H.; Xu, H.; Tian, X.; Jiang, J.; Ma, J. Image fusion meets deep learning: A survey and perspective. *Inf. Fusion* **2021**, *76*, 323–336. [CrossRef]
22. Meng, T.; Jing, X.; Yan, Z.; Pedrycz, W. A survey on machine learning for data fusion. *Inf. Fusion* **2020**, *57*, 115–129. [CrossRef]
23. Gandhi, A.; Adhvaryu, K.; Poria, S.; Cambria, E.; Hussain, A. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Inf. Fusion* **2023**, *91*, 424–444. [CrossRef]
24. Li, H.; Wu, X.-J. DenseFuse: A fusion approach to infrared and visible images. *IEEE Trans. Image Process.* **2018**, *28*, 2614–2623. [CrossRef]
25. Li, H.; Wu, X.-J.; Kittler, J. RFN-Nest: An end-to-end residual fusion network for infrared and visible images. *Inf. Fusion* **2021**, *73*, 72–86. [CrossRef]
26. Xu, H.; Ma, J.; Jiang, J.; Guo, X.; Ling, H. U2Fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 502–518. [CrossRef]
27. He, J.; Luo, X.; Zhang, Z.; Wu, X.-j. MemoryFusion: A novel architecture for infrared and visible image fusion based on memory unit. *Pattern Recognit.* **2026**, *170*, 112004. [CrossRef]
28. Ma, J.; Tang, L.; Fan, F.; Huang, J.; Mei, X.; Ma, Y. SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA J. Autom. Sin.* **2022**, *9*, 1200–1217. [CrossRef]
29. He, D.; Li, W.; Wang, G.; Huang, Y.; Liu, S. MMIF-INet: Multimodal medical image fusion by invertible network. *Inf. Fusion* **2025**, *114*, 102666. [CrossRef]
30. Dinh, L.; Krueger, D.; Bengio, Y. Nice: Non-linear independent components estimation. *arXiv* **2014**, arXiv:1410.8516.
31. Dinh, L.; Sohl-Dickstein, J.; Bengio, S. Density estimation using Real NVP. *arXiv* **2016**, arXiv:1605.08803.

32. Chen, J.; Frey, E.C.; Du, Y. Unsupervised learning of diffeomorphic image registration via transmorph. In Proceedings of the International Workshop on Biomedical Image Registration, Munich, Germany, 10–12 July 2022; pp. 96–102.
33. Kang, M.; Hu, X.; Huang, W.; Scott, M.R.; Reyes, M. Dual-stream pyramid registration network. *Med. Image Anal.* **2022**, *78*, 102379. [CrossRef] [PubMed]
34. Wu, Y.; Zhang, Y.; Ma, W.; Gong, M.; Fan, X.; Zhang, M.; Qin, A.K.; Miao, Q. RORNet: Partial-to-partial registration network with reliable overlapping representations. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *35*, 15453–15466. [CrossRef]
35. Xu, H.; Yuan, J.; Ma, J. Murf: Mutually reinforcing multi-modal image registration and fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12148–12166. [CrossRef]
36. Ma, T.; Zhang, S.; Li, J.; Wen, Y. Iirp-net: Iterative inference residual pyramid network for enhanced image registration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 11546–11555.
37. Chang, Y.; Li, Z.; Xu, W. CGnet: A correlation-guided registration network for unsupervised deformable image registration. *IEEE Trans. Med. Imaging* **2024**, *44*, 1468–1479. [CrossRef] [PubMed]
38. Velesaca, H.O.; Bastidas, G.; Rouhani, M.; Sappa, A.D. Multimodal image registration techniques: A comprehensive survey. *Multimed. Tools Appl.* **2024**, *83*, 63919–63947. [CrossRef]
39. Ghahremani, M.; Khateri, M.; Jian, B.; Wiestler, B.; Adeli, E.; Wachinger, C. H-vit: A hierarchical vision transformer for deformable image registration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 11513–11523.
40. Yang, X.; Kwitt, R.; Styner, M.; Niethammer, M. Quicksilver: Fast predictive image registration–a deep learning approach. *NeuroImage* **2017**, *158*, 378–396. [CrossRef] [PubMed]
41. Dosovitskiy, A.; Fischer, P.; Ilg, E.; Hausser, P.; Hazirbas, C.; Golkov, V.; Van Der Smagt, P.; Cremers, D.; Brox, T. Flownet: Learning optical flow with convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2758–2766.
42. Sun, D.; Yang, X.; Liu, M.-Y.; Kautz, J. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8934–8943.
43. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 764–773.
44. Dong, W.; Zhu, H.; Lin, S.; Luo, X.; Shen, Y.; Guo, G.; Zhang, B. Fusion-mamba for cross-modality object detection. *IEEE Trans. Multimed.* **2025**, *27*, 7392–7406.
45. Yang, B.; Jiang, Z.; Pan, D.; Yu, H.; Gui, G.; Gui, W. LFDT-Fusion: A latent feature-guided diffusion Transformer model for general image fusion. *Inf. Fusion* **2025**, *113*, 102639. [CrossRef]
46. Zhang, H.; Ma, J. SDNet: A versatile squeeze-and-decomposition network for real-time image fusion. *Int. J. Comput. Vis.* **2021**, *129*, 2761–2785. [CrossRef]
47. Zhao, Z.; Bai, H.; Zhu, Y.; Zhang, J.; Xu, S.; Zhang, Y.; Zhang, K.; Meng, D.; Timofte, R.; Van Gool, L. DDFM: Denoising diffusion model for multi-modality image fusion. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 8082–8093.