Multimodal Image Matching Based on Cross-Modality Completion Pre-training

Meng Yang[1] , **Fan Fan**[*1] , **Jun Huang**[1] , **Yong Ma**[1] , **Xiaoguang Mei**[1] ,
Zhanchuan Cai[2] , **Jiayi Ma**[1]
[1]Wuhan University
[2]Macau University of Science and Technology
{2024102120059, fanfan, junhwong, meixiaoguang, jiayima}@whu.edu.cn, zccai@must.edu.mo

Abstract

The differences in imaging devices cause multimodal images to have modal differences and geometric distortions, complicating the matching task. Deep learning-based matching methods struggle with multimodal images due to the lack of large annotated multimodal datasets. To address these challenges, we propose XCP-Match based on cross-modality completion pre-training. XCP-Match has two phases. (1) Self-supervised cross-modality completion pre-training based on real multimodal image dataset. We develop a novel pre-training model to learn cross-modal semantic features. The pre-training uses masked image modeling method for cross-modality completion, and introduces an attention-weighted contrastive loss to emphasize matching in overlapping areas. (2) Supervised fine-tuning for multimodal image matching based on the augmented MegaDepth dataset. XCP-Match constructs a complete matching framework to overcome geometric distortions and achieve precise matching. Two-phase training encourages the model to learn deep cross-modal semantic information, improving adaptation to modal differences without needing large annotated datasets. Experiments demonstrate that XCP-Match outperforms existing algorithms on public datasets.

1 Introduction

Multimodal images, from different sensors like visible and infrared images, provide a more comprehensive understanding of the scene than single-modal images. They are valuable for advanced vision tasks, such as image fusion [Tang *et al.*, 2022a; Ma *et al.*, 2022], object detection and tracking [Zhao *et al.*, 2023], and 3D reconstruction [Jiang *et al.*, 2021]. However, multimodal images suffer from geometric distortions such as scale variations, rotations, and perspective distortions due to differences in imaging devices, which makes it hard for computers to analyze. Therefore, multimodal image matching is needed to establish the correspondence of feature points or regions between different modal images.

Multimodal images have significant modal differences in radiometric properties. For example, visible images capture the reflected light of object, while infrared images capture the thermal radiation [Tang *et al.*, 2022b; Li *et al.*, 2013]. This leads to inconsistencies in texture, contrast, and intensity of the images. The modal differences reduce the accuracy of feature extraction, and increase the matching difficulty. Current deep learning-based matching methods train their models on rich single-modal datasets and struggle to generalize to multimodal image scenarios, mainly due to the lack of large-scale annotated datasets. These problems limit the practical application of multimodal image matching methods.

To overcome these problems, we propose XCP-Match, a multimodal image matching algorithm based on cross-modality completion pre-training. The training of XCP-Match has two phases. The first is the self-supervised pre-training phase. We develop a novel pre-training model using vision transformer (ViT) [Dosovitskiy *et al.*, 2021], VGG [Simonyan and Zisserman, 2015], cross-attention fusion module and image reconstruction module. The pre-trained model has two branches to adaptively extract features from each input image. The task of this model is to perform cross-modality completion using a new masked image modeling method (MIM), which is reconstructing the masked parts of one modal image using the visible information in another modal image (the reference image). Pre-training forces the model to learn richer and in-depth semantic features across different modalities by cross-modality completion, which enhances the model's understanding of image content, leading to more accurate matching. Pre-training is performed on a real multimodal image dataset, where the image pairs are not fully aligned. To encourage the ViT encoder to focus on extracting features from the overlapping regions, we propose the attention-weighted contrast loss. The second is the supervised fine-tuning phase, which fine-tune the overall matching network using the MegaDepth dataset [Li and Snavely, 2018], based on the pre-training weights. The MegaDepth dataset is augmented to enhance the model's adaptability to modal differences. We construct a complete matching framework for XCP-Match, including multimodal and multiscale feature extraction module, coarse-level matching module, fine-level matching module, and subpixel refinement module, to achieve precise matching. The two-phase training encourages the model to learn deep cross-modal semantic features to

[*]Corresponding author.

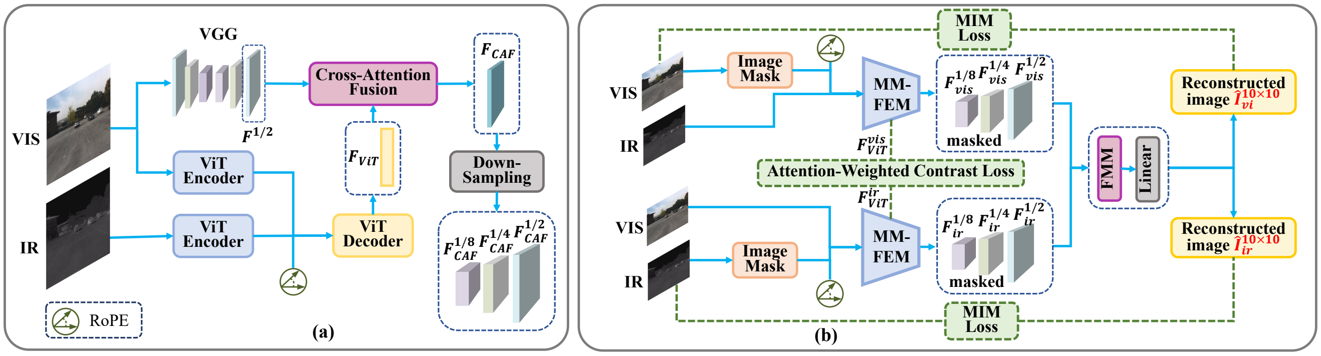
VGG
VIS
IR
$F^{1/2}$
Cross-Attention Fusion
F_{CAF}
F_{ViT}
ViT Encoder
ViT Encoder
ViT Decoder
Down-Sampling
$F^{1/8}_{CAF}$ $F^{1/4}_{CAF}$ $F^{1/2}_{CAF}$
RoPE
(a)
MIM Loss
VIS
IR
Image Mask
MM-FEM
$F^{1/8}_{vis}$ $F^{1/4}_{vis}$ $F^{1/2}_{vis}$
masked
F^{vis}_{ViT}
Attention-Weighted Contrast Loss
F^{ir}_{ViT}
VIS
IR
Image Mask
MM-FEM
$F^{1/8}_{ir}$ $F^{1/4}_{ir}$ $F^{1/2}_{ir}$
masked
FMM
Linear
Reconstructed image $\hat{I}^{10\times10}_{vi}$
Reconstructed image $\hat{I}^{10\times10}_{ir}$
MIM Loss
(b)

Figure 1: Overview of multimodal and multiscale feature extraction module and pre-training model. (a) Schematic of the visible image branch which extracts 1/8, 1/4 and 1/2 scale features of the visible image by ViT encoder, VGG, ViT decoder and cross-attention fusion module. (b) The pre-training model with two branches. In pre-training, the image reconstruction module includes FMM and linear layers.

improve adaptation to modal differences and does not require large annotated datasets. We evaluate XCP-Match on the public datasets, comparing it to state-of-the-art algorithms, and achieve the best performance.

In summary, this paper has the following contributions:

- We design a novel dual-branch pre-training model, and introduce unsupervised pre-training using a new MIM method to enhance the model's ability to extract multimodal image features.
- We design a novel attention-weighted contrast loss in pre-training to make the feature extraction focus on overlapping regions and overcome the problem of unaligned images in the dataset.
- We design a complete matching network suited for multimodal images and perform supervised fine-tuning with the augmented MegaDepth dataset for accurate optimization of matching.

2 Related Works

2.1 Feature Detector-based Matching Method

Feature detector-based matching methods involve three stages: features detecting, describing, and matching [Lowe, 2004; Bay *et al.*, 2006; Li *et al.*, 2024]. To overcome the nonlinear radiance differences (NRD) [Ye *et al.*, 2022; Li *et al.*, 2023a] of multimodal images, researchers have focused on improving the performance of feature detectors and descriptors, developing algorithms such as RIFT [Li *et al.*, 2019] and RA-MMIR [Qiu *et al.*, 2024]. Deep learning methods have also made some progress such as ReDFeat [Deng and Ma, 2023], which decouples detection from matching, and XPoint [Yagmur *et al.*, 2024], using a pre-trained encoder for feature extraction while applying geometric constraints. SemLA [Xie *et al.*, 2023] focuses on aligning key scene regions. However, these methods struggle to detect reliable keypoints in the scenario of lacking prominent features, which reduces the accuracy of subsequent matching.

2.2 Detector-Free Matching Method

Detector-free methods aim for pixel-level correspondences directly in input images, without using feature detector. These methods uses transformers' capability [Vaswani, 2017] to handle long-distance dependencies and global context, and perform more robustly in images with large geometric distortions. LoFTR [Sun *et al.*, 2021] achieves dense matching through a coarse-to-fine approach, while Efficient-LoFTR [Wang *et al.*, 2024] improves efficiency and accuracy with attention mechanisms. DualRC [Li *et al.*, 2023b] employs dual-resolution learning and neighborhood consensus. GRiD [Liu *et al.*, 2024] addresses NRD and geometric distortions by finding pixel-level matches using reference points. However, these methods require large-scale annotated datasets, which are lacking in multimodal images.

2.3 Transformation Estimation-based Method

Transformation estimation methods include flow and homography estimation. Flow estimation methods estimate each pixel's spatial displacement for registration. Zhou et al. transfers knowledge from optical flow model to cross-modal flow estimation [Zhou *et al.*, 2022]. Homography estimation methods predict image transformations. CrossHomo [Deng *et al.*, 2024] uses multimodal homography estimation module to predict homography matrices accurately. Real-world transformations' complexity challenges these methods.

2.4 MIM in Image Matching

Inspired by BERT [Kenton and Toutanova, 2019] and ViT, self-supervised MIM methods have arisen in computer vision, training models to recover complete signals from unmasked data parts. MIM methods like MAE [He *et al.*, 2022], MultiMAE [Bachmann *et al.*, 2022], and MST [Li *et al.*, 2021] excel in classification and also make breakthroughs in geometric tasks. CroCo [Weinzaepfel *et al.*, 2022] reconstructs the masked image using another view of the same scene, suitable for geometric tasks. CroCo v2 [Weinzaepfel *et al.*, 2023] builds on CroCo with enhancements. PMatch [Zhu and Liu, 2023] extends MIM to pairwise reconstruction, optimizing both encoder and decoder. Tuzcuoğlu improves LoFTR based on PMatch to match infrared and visible images [Tuzcuoğlu *et al.*, 2024]. These developments highlight MIM's role in enhancing feature representation and narrowing the domain gap between pre-training and downstream tasks.

$F_{vis/ir}$	Features processed by ViT encoder
F_{ViT}	Features of F_{vis} and F_{ir} fused by ViT decoder
F_{VGG}	Feature of I_{vis} processed by VGG
M	Normalized attention score matrix
F_{CAF}	Enhanced feature map by CAF
$F_*^{1/2,1/4,1/8}$	Feature map at $1/2$, $1/4$ and $1/8$ resolution
S	The similarity matrix between $\{F_{vis}^{1/8}, F_{ir}^{1/8}\}$
$P_{k\in(0,1)}, P^f$	The coarse-level and fine-level matching probability matrix
M_c, M_f	The final coarse-level and fine-level matching set
S^f	The similarity matrix between $\{\hat{f}_{vis}^{5\times5}, \hat{f}_{ir}^{5\times5}\}$
θ_c, θ_f	The threshold for coarse-level and fine-level matching
$\{\delta_{vis}, \delta_{ir}\}$	The local subpixel offsets for each match
$C_*, \hat{C}_*$	$(\hat{i}, \hat{j})$ coordinates before and after subpixel refinement

Table 1: List of Symbols.

3 Methodology

Our method has four modules: multimodal and multiscale feature extraction module, coarse-level and fine-level matching module, and sub-pixel refinement module. We introduce a two-phase training strategy including self-supervised pre-training and supervised fine-tuning for matching. The meaning of the symbols is shown in Table 1.

3.1 Multimodal and Multiscale Feature Extraction Module

To address the challenges of modality and viewpoint differences in multimodal images, we propose the multimodal and multiscale feature extraction module (MM-FEM) based on VGG networks and ViT. To enhance the performance of MM-FEM in extracting multimodal image features, we introduce a novel MIM method to pre-train MM-FEM.

Feature extraction: Given two multimodal images from the same scene, e.g., infrared image I_{ir} and visible image I_{vis}, they are gridded into $N = n^2$ non-overlapping patches and labeled as tokens for input to the MM-FEM. The MM-FEM has two different asymmetric branches as shown in Figure 1. The inputs to the first branch are tokens from I_{vis} and I_{ir}, respectively. The tokens of I_{vis} are fed to the VGG and the ViT encoder, respectively, while the tokens of I_{ir} are fed only to the ViT encoder with shared weights. The ViT decoder fuses the features $F_{vis} = Encoder(I_{vis})$ and $F_{ir} = Encoder(I_{ir})$ from the encoders to enable the deep interaction of multimodal image information, resulting in a cross-modal feature map of size $[B, N, C]$:

$$F_{ViT} = Decoder(F_{vis}, F_{ir}). \tag{1}$$

A multibranch VGG with residual connections processes tokens from I_{vis} and generates the base feature map of size $[B, C, H/2, W/2]$: $F_{VGG} = VGG(I_{vis})$, where H and W are the height and width of the original image, respectively.

Cross-attention fusion: To enhance the fusion of information from another modal image, we introduce a multi-headed cross-attention fusion (CAF) module to adaptively update the base feature map F_{VGG}, which enables the model to learn the richer and more robust feature representations.

In CAF, we use F_{ViT} as key and value, and use F_{VGG} as query. First, reshape F_{VGG} to size $[B, (H * W)/4, C]$. The normalized attention score matrix M is computing by the following:

$$M = Softmax(F_{VGG} \cdot F_{ViT}^T). \tag{2}$$

The enhanced feature for each location is computed by weighting and summing F_{ViT} using the attention score matrix M. Since M varies with different modal inputs, it exhibits adaptability to new multimodal images. Finally, the updated result is linearly transformed by a fully connected layer to yield the enhanced feature F_{CAF} with the same shape of F_{VGG}:

$$F_{CAF} = FC(M \cdot F_{ViT}), \tag{3}$$

where F_{CAF} is the enhanced feature map. For brevity, we omit some reshaping operations in equation.

Down-sampling: Input F_{CAF} into the VGG with different down-sampling scales to further extract the features of $1/4$ and $1/8$ original image resolution for the later coarse-to-fine matching.

As shown in Figure 1 (b), the structure of the second branch is same as the first branch. However, in the second branch, the tokens of I_{ir} are fed to both the VGG and ViT. Additionally, we use rotational position embedding (RoPE) on the image tokes before the encoder in two branches, which injects the relative position information of features.

3.2 Coarse-level Matching Module

The coarse-level matching module (CMM) uses the coarse-level feature map $F_{vis}^{1/8}$ and $F_{ir}^{1/8}$ of 1/8 resolution from MM-FEM to predict the matches at 1/8 scale.

Coarse-level feature interaction: We use the linear self-attention and cross-attention in LoFTR to interact $F_{vis}^{1/8}$ and $F_{ir}^{1/8}$, outputting $\hat{F}_{vis}^{1/8}$ and $\hat{F}_{ir}^{1/8}$, which captures the global dependencies between multimodal images.

Similarity matrix computation: Given $\hat{F}_{vis}^{1/8}$ and $\hat{F}_{ir}^{1/8}$, the similarity matrix S is computed by the following:

$$S(i,j) = \frac{1}{\gamma} \cdot \left\langle Linear(\hat{F}_{vis}^{1/8}), Linear(\hat{F}_{ir}^{1/8}) \right\rangle, \tag{4}$$

where i and j are the indexes in $\hat{F}_{vis}^{1/8}$ and $\hat{F}_{ir}^{1/8}$, respectively. $Linear(\cdot)$ denotes the linear layer, and $\langle \cdot, \cdot \rangle$ denotes the inner product. γ is the temperature parameter.

Matching probability matrix computation: Apply softmax operation to $S(i,j)$ to obtain the matching probability matrix $P_{k\in(0,1)}(i,j) = soft\max\left(S(i,\cdot)\right)_j$. P_0 and P_1 are the matching probability matrices derived from softmax along the first and zeroth dimensions, respectively.

Coarse-level match acquisition: We use the threshold θ_c to filter out high-confidence elements and select the largest as the matching pair $(\tilde{i}, \tilde{j})$:

$$\begin{aligned} M_k =& \{(\tilde{i},\tilde{j}) | P_k(\tilde{i},\tilde{j}) > \theta_c, \\ & P_k(\tilde{i},\tilde{j}) = \max_x P_k(\tilde{i},x) \; or \max_x P_k(x,\tilde{j})\}_{k\in(0,1)}. \end{aligned} \tag{5}$$

The final coarse-level matching set is $M_c = M_0 \cup M_1$.

3.3 Fine-level Matching Module

Based on M_c, the fine-level matching module (FMM) uses $F^{1/2}$ and $F^{1/4}$ to seek fine-level matches at $1/2$ resolution to improve the matching accuracy. Given the significant texture

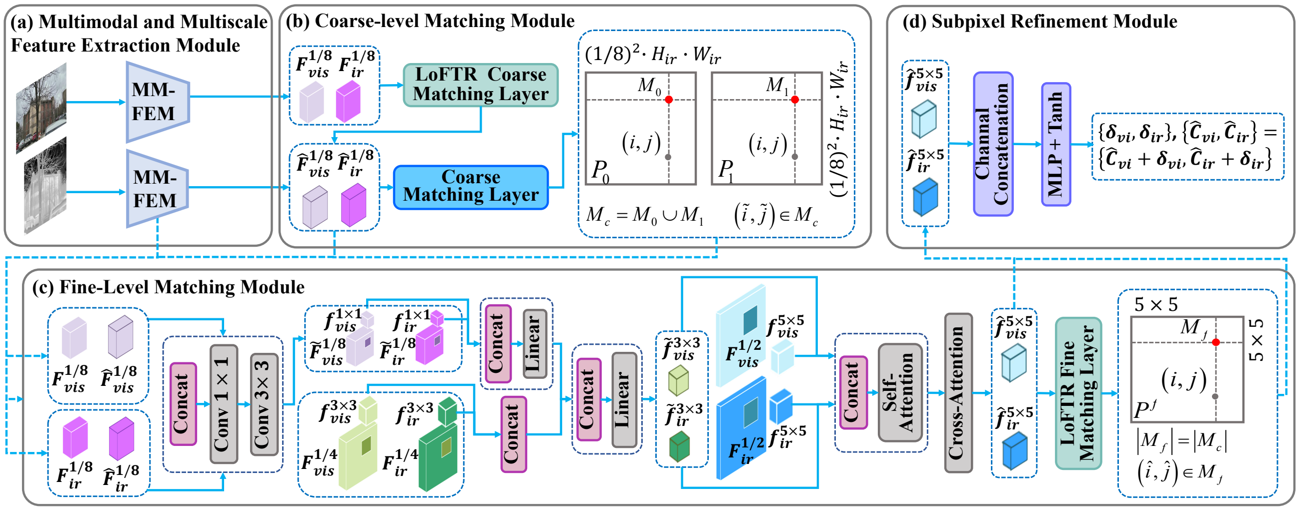
(a) Multimodal and Multiscale Feature Extraction Module
MM-FEM
MM-FEM
(b) Coarse-level Matching Module
$F_{vis}^{1/8}$ $F_{ir}^{1/8}$
LoFTR Coarse Matching Layer
$\hat{F}_{vis}^{1/8}$ $\hat{F}_{ir}^{1/8}$
Coarse Matching Layer
$(1/8)^2 \cdot H_{ir} \cdot W_{ir}$
M_0
M_1
(i,j)
P_0
P_1
$(1/8)^2 \cdot H_{ir} \cdot W_{ir}$
$M_c = M_0 \cup M_1$
$(\tilde{i},\tilde{j}) \in M_c$
(d) Subpixel Refinement Module
$\hat{f}_{vis}^{5\times5}$
$\hat{f}_{ir}^{5\times5}$
Channal Concatenation
MLP + Tanh
$\{\delta_{vi}, \delta_{ir}\}, \{\hat{C}_{vi}, \hat{C}_{ir}\} = \{C_{vi} + \delta_{vi}, C_{ir} + \delta_{ir}\}$
(c) Fine-Level Matching Module
$F_{vis}^{1/8}$ $\hat{F}_{vis}^{1/8}$
$F_{ir}^{1/8}$ $\hat{F}_{ir}^{1/8}$
Concat
Conv 1 × 1
Conv 3 × 3
$f_{vis}^{1\times1}$ $f_{ir}^{1\times1}$
$\tilde{F}_{vis}^{1/8}$ $\tilde{F}_{ir}^{1/8}$
Concat
Linear
$f_{vis}^{3\times3}$ $f_{ir}^{3\times3}$
$F_{vis}^{1/4}$ $F_{ir}^{1/4}$
Concat
Concat
Linear
$\tilde{f}_{vis}^{3\times3}$
$\tilde{f}_{ir}^{3\times3}$
$F_{vis}^{1/2}$ $f_{vis}^{5\times5}$
$F_{ir}^{1/2}$ $f_{ir}^{5\times5}$
Concat
Self-Attention
Cross-Attention
$\hat{f}_{vis}^{5\times5}$
$\hat{f}_{ir}^{5\times5}$
LoFTR Fine Matching Layer
5 × 5
5 × 5
M_f
(i,j)
P^f
$|M_f| = |M_c|$
$(\hat{i},\hat{j}) \in M_f$

Figure 2: Overview of the XCP-Match framework. (a) is the multimodal and multiscale feature extraction module. (b) is the coarse-level matching module, which predicts coarse-level matches at the 1/8 scale. (c) is the fine-level matching module, which uses 1/2 and 1/4 scale features based on the coarse-level matches to predict fine-level matches. (d) is the subpixel refinement module, which refines fine matches at the subpixel level through the regression mechanism.

differences between images of different modalities, FMM re-evaluates coarse-level matches on a finer scale, capturing detailed texture information more accurately. In addition, FMM is also part of the image reconstruction module in pre-training because it can capture texture and structural information at different scales.

Feature preprocessing: To improve the interaction of $F^{1/2}$ and $F^{1/8}$, we efficiently preprocess them. First, we cascade $\hat{F}_{vis}^{1/8}$ and $F_{vis}^{1/8}$ along the channel dimension, and apply point-by-point convolution and deep convolution with 3×3 kernel to obtain $\tilde{F}_{vis}^{1/8} = Conv_3(Conv_1(\hat{F}_{vis}^{1/8}|F_{vis}^{1/8}))$ with the same channel size as $F_{vis}^{1/4}$. Similarly, the same operation is used for $\hat{F}_{ir}^{1/8}$ and $F_{ir}^{1/8}$ to obtain $\tilde{F}_{ir}^{1/8}$.

Local window extraction: The local window pair $\{f_{vis}^{1\times1}, f_{ir}^{1\times1}\}$, $\{f_{vis}^{3\times3}, f_{ir}^{3\times3}\}$, and $\{f_{vis}^{5\times5}, f_{ir}^{5\times5}\}$ of each $(\tilde{i},\tilde{j})$ are extracted from $\left\{\tilde{F}_{vis}^{1/8}, \tilde{F}_{ir}^{1/8}\right\}$, $\left\{F_{vis}^{1/4}, F_{ir}^{1/4}\right\}$, and $\left\{F_{vis}^{1/2}, F_{ir}^{1/2}\right\}$ using window sizes of 1×1, 3×3, and 5×5, respectively. We add the absolute position bias to each window before sending it to the next layer.

Multiscale and multimodal feature passing: To pass information between $\{f_{vis}^{1\times1}, f_{ir}^{1\times1}\}$, $\{f_{vis}^{3\times3}, f_{ir}^{3\times3}\}$ and $\{f_{vis}^{5\times5}, f_{ir}^{5\times5}\}$, we perform a series of concatenation, linear layer, self-attention, cross-attention and splitting operations to obtain $\{\hat{f}_{vis}^{5\times5}, \hat{f}_{ir}^{5\times5}\}$.

Fine-level matching: For each $(\tilde{i},\tilde{j})$, compute the similarity matrix S^f between $\{\hat{f}_{vis}^{5\times5}, \hat{f}_{ir}^{5\times5}\}$, and apply a double softmax to obtain the fine-level match probability matrix P^f:

$$P^f(i,j) = soft\max\left(S^f(i,\cdot)\right)_j \cdot soft\max\left(S^f(\cdot,j)\right)_i. \quad (6)$$

Finally, the matching pair $(\hat{i},\hat{j})$ with $P^f(i,j)$ greater than the threshold θ_f and all other elements in each $(\tilde{i},\tilde{j})$ is selected as the fine-level match M_f.

3.4 Subpixel Refinement Module

The goal of this module is to refine the fine-level matches to subpixel accuracy. We concate $\{\hat{f}_{vis}^{5\times5}, \hat{f}_{ir}^{5\times5}\}$ at the fine-level match $(\hat{i},\hat{j})$, and use the MLP layer and the Tanh function to jointly predict the local subpixel offsets for each match: $\{\delta_{vis}, \delta_{ir}\} = \mathrm{Tanh}(MLP(\hat{f}_{vis}^{5\times5}|\hat{f}_{ir}^{5\times5}))$. Finally, these offsets are added on the coordinates of $(\hat{i},\hat{j})$ to obtain the subpixel-level matches:

$$\{\hat{C}_{vis}, \hat{C}_{ir}\} = \{C_{vis} + \delta_{vis}, C_{ir} + \delta_{ir}\}, \quad (7)$$

where $\{C_{vis}, C_{ir}\}$ is the coordinate of $(\hat{i},\hat{j})$ before subpixel refinement and $\{\hat{C}_{vis}, \hat{C}_{ir}\}$ is the coordinate after refinement.

3.5 Self-Supervised Pre-training

MIM method: XCP-Match is pre-trained with the pre-training model and MIM method. The goal of pre-training is to train the model to complete the masked parts of the image using another modal image as reference. The pre-training model is shown in Figure 1 (b). The inputs to first branch are the tokens of masked I_{vis} and I_{ir}. The tokens of masked I_{vis} are fed to the VGG and the ViT encoder, while the tokens of I_{ir} are fed only to the ViT encoder. The inputs to second branch are the tokens of masked I_{ir} and I_{vis}. The feature map after MM-FEM processing is directly output to FFM. Since FFM uses feature map at different scales, we mask the feature map at each scale. To mask the input images, we randomly generate 64×64 masks using the binary method to cover 50% of the image. The remaining scales of feature map use the same masking method and rate, with only the masked patch size adjusted according to the scale ratio. To make the masked areas learnable, we replace them with CNN-processed learnable masked tokens at the corresponding positions.

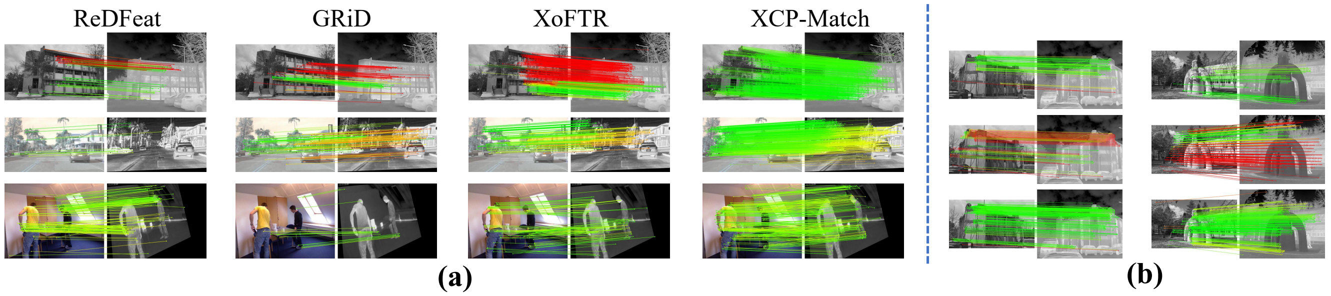

Figure 3: Qualitative results of matching line diagrams. (a) Comparison experimental results for ReDFeat, GRiD, XoFTR, and XCP-Match algorithms (left to right), using images from the METU-VisTIR, RoadScene, and TriModalHuman datasets (top to bottom). (b) Line 1: matches using multi-scale feature extraction module (MFEM) without the dual-branch architecture (DBA). Line 2: matches using DBA without MFEM (using a single-scale extractor). Line 3: matches using DBA with MFEM.

Image reconstruction module: The masked feature map is processed by FMM to obtain $\{\hat{f}_{vis}^{5\times5}, \hat{f}_{ir}^{5\times5}\}$. To reconstruct the image, we designed linear layers to upsample $\{\hat{f}_{vis}^{5\times5}, \hat{f}_{ir}^{5\times5}\}$ to the original image resolution: $\{\hat{I}_{vis}^{10\times10}, \hat{I}_{ir}^{10\times10}\} = Linear(\hat{f}_{vis}^{5\times5}, \hat{f}_{ir}^{5\times5})$. The 10×10 image window of output is used to reconstruct the masked tokens of input at the same locations.

MIM loss function: To ensure that the output image is similar to the target image, the mean square error (MSE) between the target image window $I_i^{10\times10}$ and $\hat{I}_i^{10\times10}$ is used as the similarity loss:

$$L_s = \frac{1}{n}\sum_{i=1}^{N}(\hat{I}_i^{10\times10} - I_i^{10\times10})^2. \tag{8}$$

Attention-weighted contrast loss: The image pairs in the pre-training dataset are not perfectly aligned. To encourage the ViT encoder to focus on overlapping regions, we propose the attention-weighted contrast loss based on the contrastive learning. It pulls together the feature representations of different modalities in overlapping regions, while pushing away them in non-overlapping regions. This loss facilitates MMFEM to establish the common semantic space across different modalities.

The similarity matrix between the ViT encoder output features F_{vis} and F_{ir} in the first branch is first calculated using matrix multiplication. Attention weights are derived by normalizing the similarity matrix with the softmax function, which are applied as the weighted average on F_{vis} to obtain the attention feature $\bar{F}_{vis}$. The attentional weighted contrast loss is given by the following:

$$L_{ac}^{vis} = -\frac{1}{BN}\sum_{b=1}^{BN}\log\left(\frac{\exp(sim(F_{ir}, \bar{F}_{vis})/T)}{\sum_{k=1}^{2N}\exp(sim_k/T)}\right), \tag{9}$$

where B is the batch size, N is the number of image tokens. T is the temperature coefficient. $sim(F_{ir}, \bar{F}_{vis})$ is the positive sample similarity and sim_k is the similarity of all positive and negative sample pairs. Similarly, L_{ac}^{ir} can be obtained from the second branch.

Total pre-training loss is: $L_{pretrain} = \lambda_s L_s + \lambda_{ac}^{vis} L_{ac}^{vis} + \lambda_{ac}^{ir} L_{ac}^{ir}$.

3.6 Supervised Fine-tuning

Based on the self-supervised pre-training weights, we perform supervised fine-tuning of the overall network using the dataset with ground-truth (GT). There are three loss functions for the fine-tuning.

Coarse-level matching loss: we use focus loss (FL) to supervise the matching probability matrix $P_{k\in(0,1)}$ in CMM:

$$L_c = \alpha \cdot FL(P_0, \hat{P}_0) + \beta \cdot FL(P_1, \hat{P}_1), \tag{10}$$

where $\hat{P}_0$ and $\hat{P}_1$ are the GT matching matrices for coarse matching. α and β are the weights that balance the two loss terms.

Fine-level matching loss: We design the fine-level matching loss to supervise P^f in FMM:

$$L_f = \frac{1}{M_c}\sum_{(\hat{i},\hat{j})\in M_c} FL(P_{\hat{i},\hat{j}}^f, \hat{P}_{\hat{i},\hat{j}}^f), \tag{11}$$

where $\hat{P}_{\hat{i},\hat{j}}^f$ is the GT fine-level matching matrix for $(\hat{i}, \hat{j})$.

Subpixel refinement loss: Given that the predicted matched point pairs have homogeneous coordinates $(\hat{x}_{vi}, \hat{x}_{ir})$ in the normalized coordinate system, the subpixel refinement loss can be computed using the symmetric polar distance function:

$$L_{sub} = \frac{1}{|M_c|}\sum_{(\hat{x}_{vi},\hat{x}_{ir})}\left\|\hat{x}_{vi}^T E \hat{x}_{ir}\right\|^2\left(\frac{1}{\|E^T\hat{x}_{vi}\|_{0:2}^2} + \frac{1}{\|E\hat{x}_{ir}\|_{0:2}^2}\right), \tag{12}$$

where E is the GT essential matrix obtained using the camera pose. $\|v\|_{0:2}$ denotes the first two elements of the vector v.

Total fine-tuning loss is: $L_{fine} = \lambda_c L_c + \lambda_f L_f + \lambda_{sub} L_{sub}$.

4 Experiments

4.1 Implementation Details

Self-supervised pre-training: In pre-training, we use the KAIST Multispectral Pedestrian dataset [Hwang *et al.*, 2015] for training. Pre-training is conducted using the AdamW optimizer with a learning rate of 2.5×10^{-4}, a batch size of 2, a total of 15 epochs, and 30 hours of training on 2 NVIDIA GeForce RTX 4090 GPUs.

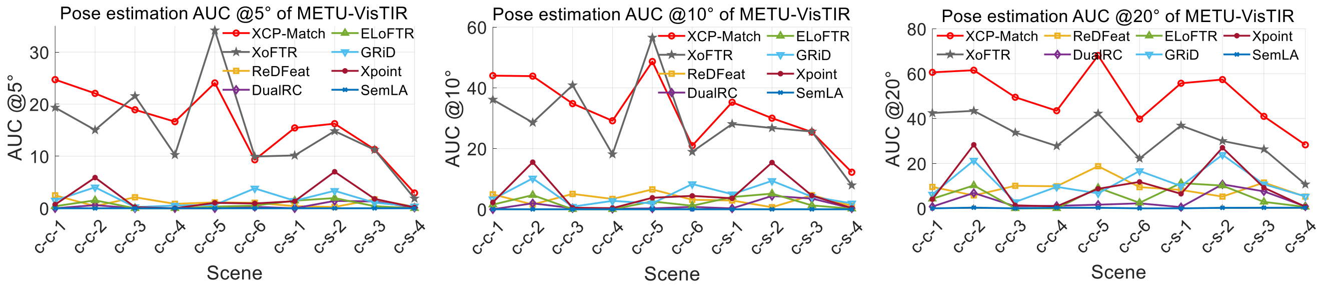

Figure 4: Visualization of the relative pose estimation results for XCP-Match and all comparison algorithms. In the x-axis, c-c and c-s denote the cloudy-cloudy and cloudy-sunny scenarios, respectively.

Method	AUC of cloud-cloud and cloud-sunny		
	@5°	@10°	@20°
DualRC	0.713/0.145	2.133/0.591	4.925/2.257
Efficient-LoFTR	0.955/0.458	2.658/1.655	6.230/4.348
ReDFeat	1.341/0.666	4.101/2.340	10.57/7.542
Xpoint	2.585/1.465	5.979/4.474	10.86/9.189
SemLA	0/0	0/0	0.212/0.156
GRiD	1.562/1.791	5.100/4.558	12.46/10.54
XoFTR	18.39/9.523	33.18/22.09	48.43/36.83
XCP-Match	**19.28/12.24**	**36.90/27.70**	**53.82/45.59**

Table 2: Quantitative results of relative pose estimation in METU-VisTIR dataset, and the values to the left and right of the '/' are the results for cloud-cloud and cloud-sunny scenarios respectively (bold fonts indicate the maximum values).

Method	AUC of RoadScene		
	@5°	@10°	@20°
DualRC	14.36	20.14	29.63
Efficient-LoFTR	13.27	19.73	31.53
ReDFeat	10.03	16.20	31.38
Xpoint	1.883	4.082	12.64
SemLA	4.910	9.837	19.56
GRiD	12.61	24.99	36.66
XoFTR	11.65	21.34	32.19
XCP-Match	**18.69**	**29.77**	**41.20**

Table 3: Quantitative results of homography estimation in the RoadScene dataset (bold fonts indicate the maximum values).

Supervised fine-tuning: In fine-tuning, we use the MegaDepth dataset [Li and Snavely, 2018]. To enhance the robustness to modal differences, we adopt the operation of XoFTR to augment data. Fine-tuning is conducted using the AdamW optimizer with a learning rate of 2.5×10^{-4}, a batch size of 2, a total of 25 epochs, and 125 hours of training on 2 NVIDIA GeForce RTX 4090 GPUs.

Hyperparameter settings: The thresholds in the matching network are set to: $\theta_c = 0.3$, $\theta_f = 0.1$. The settings in the loss function are set to: $\lambda_c = 0.5, \lambda_f = 0.3, \lambda_{sub} = 10^4$, $\lambda_s = 1, \lambda_{ac}^{vis} = \lambda_{ac}^{ir} = 0.25$.

4.2 Relative Pose Estimation

Dataset: To evaluate the performance of XCP-Match for relative pose estimation in visible-infrared image pairs, we test it on the METU-VisTIR dataset [Tuzcuoğlu *et al.*, 2024]. The dataset has thermal infrared and visible image pairs from six scenarios with GT camera poses. Four of these scenarios have both cloudy and sunny conditions, while the other two scenarios have only cloudy conditions.

Comparison scheme: XCP-Match processes the input images and generates matched point pairs. We use RANSAC [Fischler and Bolles, 1981] with a threshold of 3 to filter correct matching point pairs. During testing, the longer image side is set to 640 pixels to standardize sizes. We evaluate the methods independently on the six scenarios of the dataset to examine performance differences under different weather conditions. We use the area under curve (AUC) at 5°, 10° and 20° thresholds as evaluation metrics, measuring the maximum angular deviation from the GT in rotation and translation. We compared XCP-Match with the following publicly available methods: ReDFeat [Deng and Ma, 2023], GRiD [Liu *et al.*, 2024], Xpoint [Yagmur *et al.*, 2024], SemLA [Xie *et al.*, 2023], XoFTR [Tuzcuoğlu *et al.*, 2024], DualRC [Li *et al.*, 2023b] and Efficient-LoFTR [Wang *et al.*, 2024].

Results: As shown in Table 2 and Figure 4, XCP-Match achieves significantly higher AUC than other algorithms at all thresholds for the cloudy-cloudy and cloudy-sunny datasets. XCP-Match also achieves a significant performance advantage in most scenarios. The performance on the cloudy-sunny dataset is lower than on the cloudy-cloudy dataset, due to increased image feature variation from light and temperature differences, which makes matching and pose estimation more challenging. Figure 3 (a) illustrates the qualitative results.

4.3 Homography Transformation Estimation

Dataset: To test XCP-Match for homography estimation, we use the RoadScene dataset [Xu *et al.*, 2020b; Xu *et al.*, 2020a]. We randomly generate a unique homography matrix and apply it to the original images as GT. The homography matrices include random translations of $[-10, 10]$, random rotations of $[-10, -10]$, random scaling of $[0.8, 1.2]$, random shear angles of $[-0.1, 0.1]$, and random perspective transformations of $[-0.001, 0.001]$.

Comparison scheme: The evaluation metrics still use

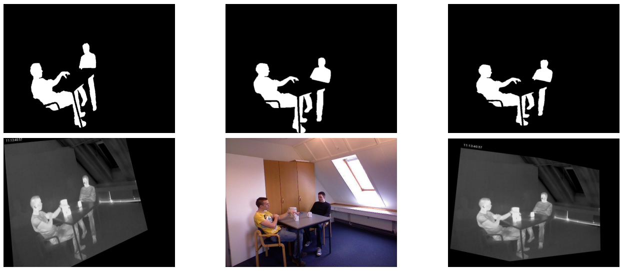

Figure 5: Qualitative results of image registration. The second row, from left to right, shows the infrared, RGB, and the registered infrared image. Images in the first row represent the segmentation annotations corresponding to each image in the second row.

RGB-FIR	Metrics	Scene1	Scene2	Scene3
DualRC	LTA	0.8231	0.7078	0.6369
	IoU	0.6263	0.4566	0.4473
Efficient-LoFTR	LTA	0.8296	0.6391	0.5393
	IoU	0.6378	0.5331	0.3362
ReDFeat	LTA	0.7343	0.6648	0.6295
	IoU	0.6209	0.3364	0.3900
Xpoint	LTA	0.7171	0.5302	0.5668
	IoU	0.6308	0.4516	0.4658
SemLA	LTA	0.7769	0.6936	0.6108
	IoU	0.7124	0.6430	0.5135
GRiD	LTA	0.8549	0.8006	0.6284
	IoU	0.7500	0.7172	0.4967
Cross-RAFT	LTA	0.8102	0.7967	0.7102
	IoU	0.7469	0.7388	0.5812
XoFTR	LTA	0.8358	0.8273	0.6883
	IoU	0.7923	0.7943	0.6062
XCP-Match	LTA	**0.8487**	**0.8389**	**0.7268**
	IoU	**0.8157**	**0.7994**	**0.6192**

Table 4: Quantitative results of image registration experiments in the three indoor scenes of the TriModalHuman dataset (bold fonts indicate the maximum values).

AUC. Since the homography matrix describes geometric transformations in planar scenes, we approximate the scene as planar to estimate the camera pose.

Results: As shown in Table 3 and Figure 3 (a), the AUC of XCP-Match are significantly higher than those of other methods at all thresholds, with the performance gap increasing as the threshold increases. Figure 3 (a) shows that XCP-Match accurately aligns the feature points between the source and target image, which ensures the transformed image maintains good geometric consistency and structural integrity, even with scale variations, perspective distortions and rotations.

4.4 Image Registration Test with Segmentation Annotations

Dataset: To test XCP-Match's performance for image registration, we evaluate it on the TriModalHuman dataset [Palmero *et al.*, 2016]. It contains 5724 RGB-Depth-FIR triples across three indoor scenes, with human body segmentation annotations. We use RGB-FIR images and select only those with distinct human segmentations for testing. To avoid causing damage to the segmentation annotations, we only apply a small degree of random homography transformation.

Comparison scheme: We estimate the homography transformation based on image matching results and apply it to the segmentation annotations. The algorithm's performance is tested by evaluating the alignment between the transformed and target segmentation annotations. The evaluation metrics are LTA (Label Transfer Accuracy) and IoU (Intersection over Union), respectively. We add Cross-RAFT [Zhou *et al.*, 2022] as the comparison algorithm.

Results: As shown in Table 4 and Figure 4, XCP-Match achieves the best performance in all three RGB-FIR indoor scenes. XCP-Match focuses on salient regions and generate denser and more homogeneous matches. Due to this characteristic, XCP-Match can better focus and align human segmentation annotations to achieve best results in LTA and IoU as shown in Figure 5.

Method	AUC of ablation study		
	@5°	@10°	@20°
w/o pre-training	10.63	25.99	43.10
w/o pre-training model	10.61	25.24	40.58
w/o attention-weighted contrast loss	11.80	26.75	43.50
w/o RoPE	11.98	26.93	43.71
XCP-Match	**12.24**	**27.70**	**45.59**

Table 5: Ablation study of XCP-Match. All experiments are performed in the cloud-sunny scenarios of the METU-VisTIR dataset (bold fonts indicate the maximum values).

4.5 Ablation Study

To verify the effectiveness of the modules and training strategy in XCP-Match, we perform the ablation experiments in the METU-VisTIR dataset. The results are shown in Table 5 and Figure 3 (b). (1) w/o pre-training: Training directly on MegaDepth without pre-training. The results show significant performance degradation, indicating the importance of pre-training for effective feature learning. (2) w/o pre-training model: Using only VGG in MM-FEM and removing other components. The results show a significant decrease in AUC, indicating its critical role in multimodal feature extraction and interaction. (3) w/o attention-weighted contrast loss: Removing this loss during pre-training leads to decrease in AUC, as this loss helps the ViT encoder focus on overlapping regions. (4) w/o RoPE: Removing the RoPE in ViT encoder. The results show that lacking position information negatively affects the algorithm's matching performance. (5) The results of Figure 3 (b) show that the absence of either component degrades matching performance, underscoring their importance for cross-modal feature learning.

5 Conclusion

In this study, our XCP-Match introduces a two-phase training strategy. In pre-training phase, we design a novel dual-branch pre-training model and MIM method to achieve cross-modality completion. The attention-weighted contrastive loss is designed to ensure the model focuses on overlapping regions. The fine-tuning phase is based on the augmented MegaDepth dataset to enhance its robustness against modal differences. XCP-Match constructs a complete matching framework. Our evaluations on public datasets demonstrate that XCP-Match outperforms state-of-the-art algorithms.

References

[Bachmann *et al.*, 2022] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, pages 348–367, 2022.

[Bay *et al.*, 2006] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European Conference on Computer Vision*, pages 404–417, 2006.

[Deng and Ma, 2023] Yuxin Deng and Jiayi Ma. ReDFeat: Recoupling detection and description for multimodal feature learning. *IEEE Transactions on Image Processing*, 32:591–602, 2023.

[Deng *et al.*, 2024] Xin Deng, Enpeng Liu, Chao Gao, Shengxi Li, Shuhang Gu, and Mai Xu. CrossHomo: Cross-modality and cross-resolution homography estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:5725–5742, 2024.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, pages 1–12, 2021.

[Fischler and Bolles, 1981] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24:381–395, 1981.

[He *et al.*, 2022] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[Hwang *et al.*, 2015] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045, 2015.

[Jiang *et al.*, 2021] Xingyu Jiang, Jiayi Ma, Guobao Xiao, Zhenfeng Shao, and Xiaojie Guo. A review of multimodal image matching: Methods and applications. *Information Fusion*, 73:22–71, 2021.

[Kenton and Toutanova, 2019] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2, 2019.

[Li and Snavely, 2018] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018.

[Li *et al.*, 2013] Zhao-Liang Li, Hua Wu, Ning Wang, Shi Qiu, José A Sobrino, Zhengming Wan, Bo-Hui Tang, and Guangjian Yan. Land surface emissivity retrieval from satellite data. *International Journal of Remote Sensing*, 34(9-10):3084–3127, 2013.

[Li *et al.*, 2019] Jiayuan Li, Qingwu Hu, and Mingyao Ai. RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform. *IEEE Transactions on Image Processing*, 29:3296–3310, 2019.

[Li *et al.*, 2021] Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, et al. MST: Masked self-supervised transformer for visual representation. *Advances in Neural Information Processing Systems*, 34:13165–13176, 2021.

[Li *et al.*, 2023a] Jiayuan Li, Qingwu Hu, and Yongjun Zhang. Multimodal image matching: A scale-invariant algorithm and an open dataset. *ISPRS Journal of Photogrammetry and Remote Sensing*, 204:77–88, 2023.

[Li *et al.*, 2023b] Xinghui Li, Kai Han, Shuda Li, and Victor Prisacariu. DualRC: A dual-resolution learning framework with neighbourhood consensus for visual correspondences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:236–249, 2023.

[Li *et al.*, 2024] Zizhuo Li, Shihua Zhang, and Jiayi Ma. U-match: Exploring hierarchy-aware local context for two-view correspondence learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:10960–10977, 2024.

[Liu *et al.*, 2024] Yuyan Liu, Wei He, and Hongyan Zhang. GRiD: Guided refinement for detector-free multimodal image matching. *IEEE Transactions on Image Processing*, 33:5892–5906, 2024.

[Lowe, 2004] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.

[Ma *et al.*, 2022] Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7):1200–1217, 2022.

[Palmero *et al.*, 2016] Cristina Palmero, Albert Clapés, Chris Bahnsen, Andreas Møgelmose, Thomas B Moeslund, and Sergio Escalera. Multi-modal rgb–depth–thermal human body segmentation. *International Journal of Computer Vision*, 118:217–239, 2016.

[Qiu *et al.*, 2024] Junhui Qiu, Hao Li, Hualong Cao, Xiangshuai Zhai, Xuedong Liu, Meng Sang, Kailong Yu, Yunpin Sun, Yang Yang, and Pan Tan. RA-MMIR: Multi-modal image registration by robust adaptive variation attention gauge field. *Information Fusion*, 105:102215, 2024.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, pages 1–10, 2015.

[Sun *et al.*, 2021] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of* *the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8922–8931, 2021.

[Tang *et al.*, 2022a] Linfeng Tang, Jiteng Yuan, and Jiayi Ma. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Information Fusion*, 82:28–42, 2022.

[Tang *et al.*, 2022b] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83:79–92, 2022.

[Tuzcuoğlu *et al.*, 2024] Önder Tuzcuoğlu, Aybora Köksal, Buğra Sofu, Sinan Kalkan, and A Aydin Alatan. XoFTR: Cross-modal feature matching transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4275–4286, 2024.

[Vaswani, 2017] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, pages 6000–6010, 2017.

[Wang *et al.*, 2024] Yifan Wang, Xingyi He, Sida Peng, Dongli Tan, and Xiaowei Zhou. Efficient LoFTR: Semi-dense local feature matching with sparse-like speed. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21666–21675, 2024.

[Weinzaepfel *et al.*, 2022] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. *Advances in Neural Information Processing Systems*, 35:3502–3516, 2022.

[Weinzaepfel *et al.*, 2023] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17969–17980, 2023.

[Xie *et al.*, 2023] Housheng Xie, Yukuan Zhang, Junhui Qiu, Xiangshuai Zhai, Xuedong Liu, Yang Yang, Shan Zhao, Yongfang Luo, and Jianbo Zhong. Semantics lead all: Towards unified image registration and fusion from a semantic perspective. *Information Fusion*, 98:101835, 2023.

[Xu *et al.*, 2020a] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2020.

[Xu *et al.*, 2020b] Han Xu, Jiayi Ma, Zhuliang Le, Junjun Jiang, and Xiaojie Guo. FusionDN: A unified densely connected network for image fusion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12484–12491, 2020.

[Yagmur *et al.*, 2024] Ismail Can Yagmur, Hasan F Ates, and Bahadir K Gunturk. Xpoint: A self-supervised visual-state-space based architecture for multispectral image registration. *arXiv preprint arXiv:2411.07430*, 2024.

[Ye *et al.*, 2022] Yuanxin Ye, Tengfeng Tang, Bai Zhu, Chao Yang, Bo Li, and Siyuan Hao. A multiscale framework with unsupervised learning for remote sensing image registration. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022.

[Zhao *et al.*, 2023] Haojie Zhao, Junsong Chen, Lijun Wang, and Huchuan Lu. Arkittrack: a new diverse dataset for tracking using mobile RGB-D data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5126–5135, 2023.

[Zhou *et al.*, 2022] Shili Zhou, Weimin Tan, and Bo Yan. Promoting single-modal optical flow network for diverse cross-modal flow estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3562–3570, 2022.

[Zhu and Liu, 2023] Shengjie Zhu and Xiaoming Liu. Pmatch: Paired masked image modeling for dense geometric matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21909–21918, 2023.