Cross-Modal Spherical Aggregation for Weakly Supervised Remote Sensing Shadow Removal

Kaichen Chi, Wei Jing, Junjie Li, Qiang Li, *Member, IEEE*, and Qi Wang, *Senior Member, IEEE*

***Abstract*—Shadows are dark areas, typically rendering low illumination intensity. Admittedly, the infrared image can provide robust illumination cues that the visible image lacks, but existing methods ignore the collaboration between heterogeneous modalities. To fill this gap, we propose a weakly supervised shadow removal network with a spherical feature space, dubbed S2-ShadowNet, to explore the best of both worlds for visible and infrared modalities. Specifically, we employ a modal translation (visible-to-infrared) model to learn the cross-domain mapping, thus generating realistic infrared samples. Then, Swin Transformer is utilized to extract strong representational visible/infrared features. Simultaneously, the extracted features are mapped to the smooth spherical manifold, which alleviates the domain shift through regularization. Well-designed similarity loss and orthogonality loss are embedded into the spherical space, prompting the separation of private visible/infrared features and the alignment of shared visible/infrared features through constraints on both representation content and orientation. Such a manner encourages implicit reciprocity between modalities, thus providing a novel insight into shadow removal. Notably, ground truth is not available in practice, thus S2-ShadowNet is trained by cropping shadow and shadow-free patches from the shadow image itself, avoiding stereotypical and strict pair data acquisition. More importantly, we contribute a large-scale weakly supervised shadow removal benchmark that makes shadow removal independent of specific scenario constraints possible. Extensive experiments demonstrate that S2-ShadowNet outperforms state-of-the-art methods in both qualitative and quantitative comparisons. The code and benchmark are available at https://github.com/chi-kaichen/S2-ShadowNet.**

***Index Terms*—Shadow removal, multi-modal vision, spherical space, weakly supervised learning.**

I. Introduction

THE shadow is a prevalent physical phenomenon in nature, typically formed when light sources are occluded [1]. Unfortunately, undesired shadows bring further complexities and challenges to subsequent vision tasks, *e.g.*, object detection [2], semantic segmentation [3], [4], and scene classification [5]. Therefore, shadow removal is a nontrivial step in computer vision processing, and has received increasing attention [6], [56]–[60].

Early shadow removal methods analyze the statistics of illumination to detect and remove shadows through prior knowledge, such as gradient [7], morphology [8], illumination condition [9], [10], and user interaction [11]. However, the handcrafted prior is fundamentally ill-posed, leading to traditional-based methods are typically unstable and sensitive when facing challenging shadow distortions.

Benefiting from the remarkable generalization capability of neural networks, deep learning-based methods lead to substantial improvement in shadow removal [12]–[14]. Pioneering works have been explored from diverse perspectives, *e.g.*, multi-task decoupling [6], [15], imagery decomposition [16], [17], and exposure fusion [18]. However, the above methods employ a set of paired shadow and shadow-free versions to train neural networks in a fully supervised manner. As is well known, capturing high-quality image pairs in uncontrolled natural environments is both difficult and expensive due to dynamic changes in illumination and surface features. To

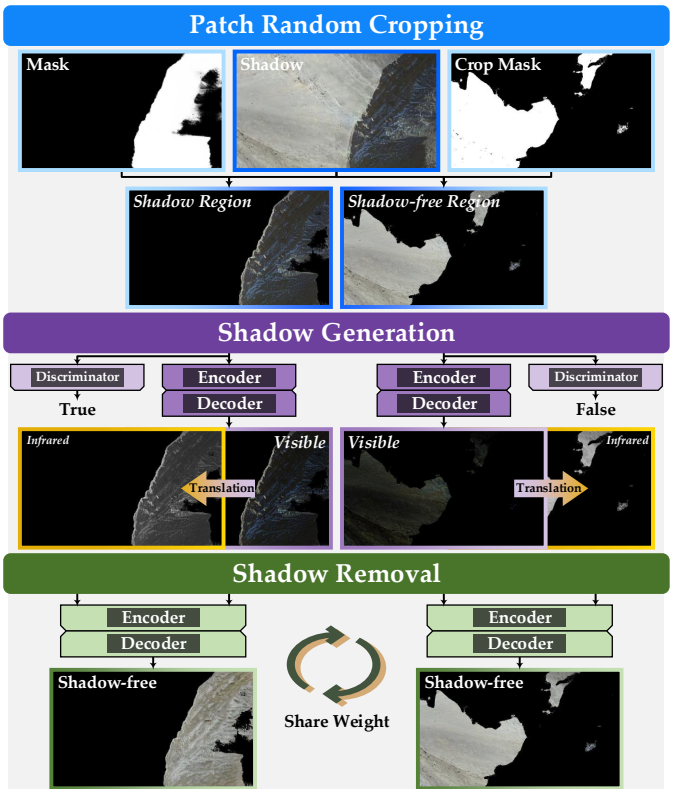

Fig. 1. The schematic illustration of our basic idea. S2-ShadowNet crops shadow and shadow-free regions from the image itself to control domain gap [23]. Shadow generation makes the elimination of the dependence on shadow-free samples possible. Besides, the introduction of complementary information from diverse modalities stimulates the potential of shadow removal.

This work was supported in part by the National Natural Science Foundation of China under Grant 62301385, 62471394, and U21B2041, and in part by the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University, PR China under Grant CX2024107. (Corresponding author: Qi Wang.)

The authors are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: chikaichen@mail.nwpu.edu.cn, wei_adam@mail.nwpu.edu.cn, junjieli@mail.nwpu.edu.cn, liqmges@gmail.com, crabwq@gmail.com).

this end, recent researches [19]–[21] concentrate on unsupervised learning to shadow removal. Unfortunately, unsupervised methods typically produce inconsistent hue and luminosity due to the huge domain gap between non-corresponding shadow and shadow-free images [22], [23]. Therefore, weakly supervised learning for cropping content-similar shadow and shadow-free regions from the same image is preferable. Such a manner is attractive in controlling domain discrepancies.

In this paper, S2-ShadowNet crops region pairs based on masking operation, coupled with a shadow generation strategy that absorbs the desirable property of dark pixels to generate pseudo shadows for shadow-free regions, thus constructing training signals that support weakly supervised learning, as depicted in Fig. 1. Subsequently, the shadow removal paradigm accomplishes revolutionary shadow recovery effects through cross-modal feature decomposition, beyond that of a single modality. Specifically, visible and infrared features are projected into a spherical space, and then under the dual constraints of orthogonality and similarity, illumination components shared between modalities are aligned while unique private features (texture and noise, *etc.*) are inversely separated. Without bells and whistles, the dependencies between diverse modalities are gracefully modeled.

In a nutshell, the main contributions can be summarized in three-fold. ❶ S2-ShadowNet is the first work that employs the multi-modal technique to conquer shadow removal, which provides heuristic insights for subsequent research. ❷ WSSR is the first dataset serving weakly supervised remote sensing shadow removal. To enhance usability, WSSR additionally provides non-consistent shadow-free samples required for unsupervised learning. Therefore, WSSR provides a qualitative and quantitative research platform for both weakly supervised and unsupervised shadow removal methods. ❸ The unparalleled performance of S2-ShadowNet on WSSR demonstrates that freedom from consistent shadow and shadow-free sample pairs acquisition is feasible. More specifically,

- **Perspective contribution.** We rethink the shadow removal task from the perspective of multi-modal collaboration, relighting shadow pixels through the dynamic update of illumination-related dominant modality. To our best knowledge, this is the first attempt to remove shadow effects by introducing infrared knowledge.
- **Technical contribution.** We propose a cross-modal spherical aggregation network to explore the interaction of visible and infrared modalities. Multi-modal feature representations are mapped to the spherical space, and the inner product between private and shared vectors is employed to drive separation and alignment, thus removing shadow traces.
- **Practical contribution.** We construct a large-scale real-world weakly supervised shadow removal benchmark (WSSR), which provides a platform for qualitative to quantitative analysis and motivates the development of deep learning-based shadow removal.

II. Related Work

In this section, we first review multi-modal image translation and multi-modal image collaboration strategies, followed by a discussion on existing shadow removal methods.

A. Multi-modal Image Translation

In view of the robustness of infrared imaging in poor illumination environments, visible-to-infrared translation schemes have been actively investigated. Lee *et al.* [24] proposed a style-controlled RGB-to-Infrared translation network, which achieves the preservation of key dynamic details by incorporating the edge-guided loss. Zhang *et al.* [25] employed a pix2pix-based translation component and explored the potential capability of frequency filters to alleviate data differ. Similarly, Li *et al.* [26] translated cityscapes under diverse lighting conditions to thermal cityscapes based on pix2pixHD, thus serving the subsequent semantic urban scenario understanding. Devaguptapu *et al.* [27] designed a visible semantics-guided pseudo multi-modal translation framework, which borrows domain adaptation knowledge for cross-modal alignment. Kniaz *et al.* [28] embedded infrared histograms and infrared feature descriptors into a conditional adversarial network to produce realistic cross-modal pedestrian and car samples. Gan *et al.* [29] encouraged a generalization-promoting translation network to focus on domain-invariant feature representations, thus reducing negative transfer from visible to thermal infrared. In summary, the booming development of multi-modal image translation technology alleviates the scarcity of infrared data, making possible multi-modal collaboration.

B. Multi-modal Image Collaboration

Multi-modal image collaboration aims to enjoy the mutual benefits between multiple sensors, thus achieving the superior performance on practical tasks. Cao *et al.* [30] contributed a multi-modal gated mixture framework, which recovers texture and contrast in low-light scenarios by integrating local experts and global experts. Zhang *et al.* [31] embedded a center-guided visible-infrared pair mining loss into cross-modality networks to alleviate the interference of significant light changes for person re-identification. Mazhar *et al.* [32] employed a Gumbel-Softmax-driven feature fusion mechanism to sample from discrete multi-modal distributions, and then stochastically fused discriminative representations to achieve nighttime vehicle detection. Xie *et al.* [33] proposed a hallucination network to integrate visible and thermal infrared information, and then used an illumination-aware loss to regress visible features from similar hallucination. Such a manner effectively promotes the robustness of nighttime pedestrian detection. Lee *et al.* [34] leveraged the depth and thermal infrared sequences as physical priors for visible sequences, recognizing facial emotions by learning spatiotemporal attention volumes. Built on the feature decomposition, Jian *et al.* [35] separated the structure and texture of visible and infrared images, and then stacked sparse encoders to construct local and global saliency maps, finally implementing nighttime saliency analysis. Obviously, the complementarity of multiple modalities is a quite promising direction, deserving to be studied in the context of shadow removal.

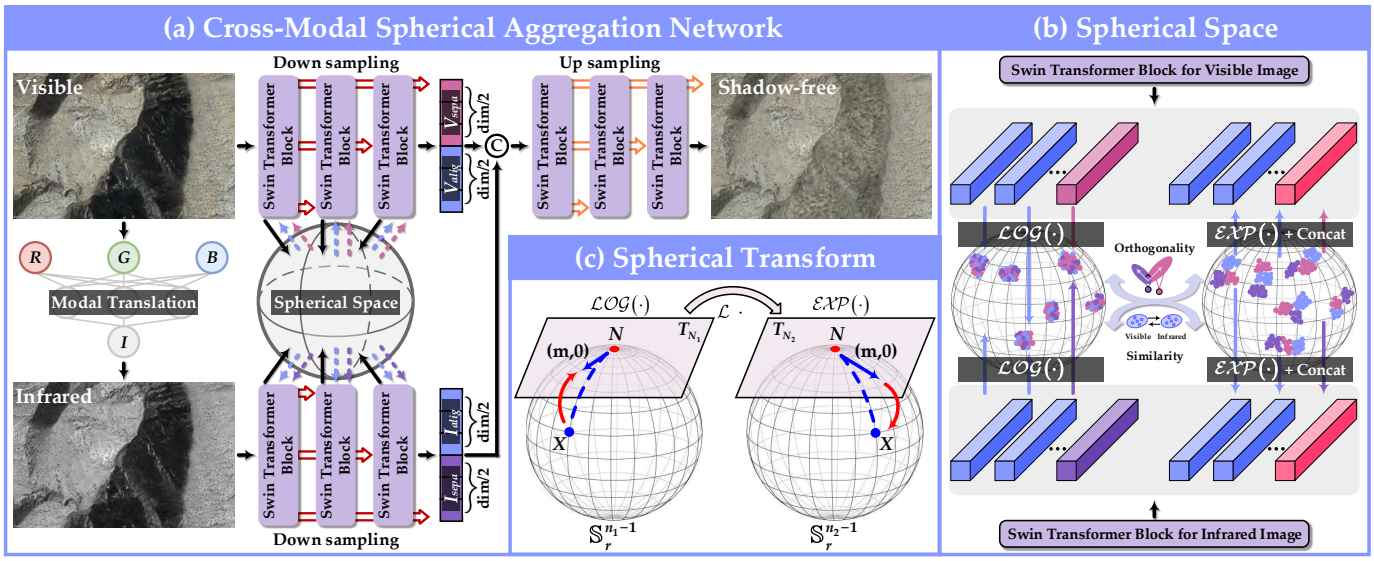

Fig. 2. (a) Overview of the architecture of S2-ShadowNet. We employ pretrained U2Fusion to generate infrared samples, and then leverage Swin Transformer to select the optimal multi-modal representation combination. Multi-modal representations are projected into the spherical space through spherical transform to accomplish decomposition, *i.e.*, alignment and separation. Under both spatial-wise and pixel-wise constraints, the full visible feature and the shared infrared feature are integrated into a shadow-free image. (b) The schematic illustration of spherical space. (c) The schematic illustration of spherical transform.

C. Shadow Removal

Traditional-based methods typically focus on exploring diverse physical shadow properties. Arbel and Hel-Or [36] employed cubic smoothing splines to calculate the per-pixel scale factor in shadows of varying width and profile, thus handling non-uniform shadows. Guo *et al.* [37] predicted the relative illumination conditions between shadow and shadow-free regions to remove shadows. Gong and Cosker [11] used the user interaction of shadow and lit regions to highlight shadow boundary intensity changes, and then accurately and robustly removed shadows by inverse scaling non-uniform field. Zhang *et al.* [38] used the texture similarity to establish the correspondence between shadow and shadow-free patches, and designed an illumination reconstruction operator to remove shadow effects. Unfortunately, traditional-based methods hardly cover the real-world shadow detection and removal due to the high dependence on specific physical priors.

With the availability of large-scale benchmarks [15], [19], deep learning-based methods achieve great breakthroughs. Cun *et al.* [39] alleviated color inconsistency and artifacts on shadow boundaries by aggregating dilated multi-contexts. Guo *et al.* [40] introduced the Retinex theory into the Transformer backbone, aiming to recover shadow regions by exploiting shadow-free regions. Wan *et al.* [41] employed a style estimator to explore the style representation of shadow-free regions, and then designed a learnable normalization to accomplish harmonious visual appearance. Liu *et al.* [42] integrated the image reconstruction, shadow matte estimation, and shadow removal branches to generate an intensive information flow to recover illumination intensity. Zhu *et al.* [43] coupled the learning processes of shadow removal and shadow synthesis in a unified framework, and then recovered colors and background contents through two-way constraints. Works [22], [23] closely related to this paper cropped windows or random regions from the same image to learn mappings between shadow and shadow-free domains. Nevertheless, these shadow removal methods neglect statistical co-occurrences between non-consistent modalities, leading to unsatisfactory robustness and generalization.

III. Methodology

The key components of S2-ShadowNet include a modal translation stage, a shadow generation stage, and a shadow removal stage. The modal translation stage employs pretrained U2Fusion [44] to provide realistic infrared signals, while the shadow generation stage guides the shadow appearance rendering through adversarial learning [23]. As a unique technical contribution, the shadow removal stage utilizes the spherical space to integrate visible path and infrared path, while the joint loss constraint serves as a tool to drive the alignment and separation of multi-modal representations, as shown in Fig. 2. In what follows, we detail the shadow removal stage, *i.e.*, the cross-modal spherical aggregation network.

A. Spherical Transform

The visible image tends to display high-frequency information, such as details. Conversely, the infrared image is insensitive to boundaries and textures, focusing on reflecting illumination through thermal radiation. Therefore, we propose to leverage the infrared modality to provide luminance information missing in the shadow region of the visible modality. Unfortunately, the cross-modal heterogeneity of both makes modality fusion tricky. To overcome this challenge, we expect to constrain the feature extraction space to enable cross-modal collaboration. For example, illumination features of infrared and visible objects can be aligned in feature space. Moreover, unique private features such as texture of visible objects and noise of infrared objects should be separated conversely. This

Algorithm 1 Code Implementation for $\mathcal{LOG}(\cdot)$.

Require: $X \in \mathbb{R}^n$, radius $r = 1$
1: Create an all-zero tensor $o \in \mathbb{R}^{n-1}$ based on X
2: $N = \mathrm{Cat}[o, r],\ N \in \mathbb{R}^n$
3: $\cos\alpha$: Calculate the cosine similarity between X and N
4: $\frac{\alpha}{\sin\alpha}$: Calculate the inverse cosine of $\cos\alpha$
5: $\mathcal{LOG}_N(X)$: $\frac{\alpha}{\sin\alpha}(X - N\cos\alpha)$
6: **return** $\mathcal{LOG}_N(X)$

Algorithm 2 Code Implementation for $\mathcal{EXP}(\cdot)$.

Require: $\hat{m} = (m, 0)$, radius $r = 1$
1: Create an all-zero tensor $o \in \mathbb{R}^{n-1}$ based on $\hat{m}$
2: $N = \mathrm{Cat}[o, r],\ N \in \mathbb{R}^n$
3: $\beta : \|\hat{m}\|/r$
4: $\mathcal{EXP}_N(\hat{m}) : N\cos\beta + \hat{m}\frac{\sin\beta}{\beta}$
5: **return** $\mathcal{EXP}_N(\hat{m})$

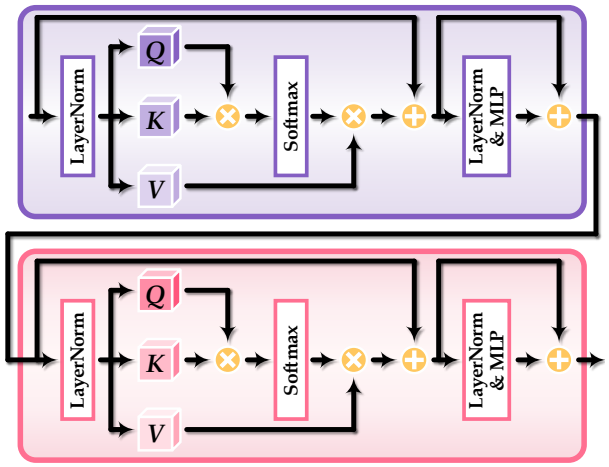

Fig. 3. The schematic illustration of the Swin Transformer block [49], which performs feature extraction through a multi-head self-attention mechanism.

means that the distance between illumination features and interference features should be pushed away [45].

Euclidean distance (*e.g.*, ℓ_1 distance or ℓ_2 distance) is often preferred for distance measurement. However, the Euclidean distance compromises the cross-modal feature since it is sensitive and tenuous to scale [45]. Fortunately, the spherical space transform provides the possibility of solving this issue, and it is gradually being applied to ship recognition [48] and super-resolution [45]. Shang *et al.* [48] employed a spherical space classifier to address large intraclass variability and small intraclass separability of ship targets. Zhao *et al.* [45] used the spherical space to aggregate shared RGB and depth features for depth map super-resolution. The spherical space is admitted because distances of spherical features are regularized, providing strong domain adaptivity [46]. Such a manner ensures that multi-modal features with different scales are elegantly aligned and separated without losing their respective unique properties [47].

Specifically, the spherical space transform is composed of an exponential mapping, a tangent space transform, and a logarithmic mapping. As depicted in Fig. 2(c), this transform process consists of projecting the feature of the spherical space $\mathbb{S}_r^{n_1-1}$ to the corresponding tangent space T_{N_1} through a spherical logarithmic mapping. Notably, T_{N_1} is essentially a hyperplane. Subsequently, the projected feature is fed through a linear transform to the tangent space T_{N_2} of the spherical space $\mathbb{S}_r^{n_2-1}$. Finally, the transformed feature is projected to the spherical space $\mathbb{S}_r^{n_2-1}$ through a spherical exponential mapping.

$$\begin{aligned} \mathcal{M} &= r(\mathcal{M}_c/\|\mathcal{M}_c\|) : \|\mathcal{M}\| = r,\ c \in \{\mathcal{V}, \mathcal{I}\} \\ \mathcal{T} &= \mathcal{EXP}_N(\mathcal{L}(\mathcal{LOG}_N(\mathcal{M}))), \end{aligned} \tag{1}$$

where $\mathcal{EXP}(\cdot)$, $\mathcal{L}(\cdot)$, and $\mathcal{LOG}(\cdot)$ represent the spherical exponential mapping, the linear transform in tangent space, and the spherical logarithmic mapping, respectively. Besides, $\mathcal{M}_c$ represents multi-modal features, $\mathcal{M}$ represents the normalized feature, N represents the north pole, and r represents the radius of spherical space.

Spherical Logarithmic Mapping. Given that the spherical space is non-Euclidean, a linear transform cannot be established between spherical spaces [48]. In contrast, the tangent space is a Euclidean. Therefore, we employ logarithmic and exponential maps as tools to bridge spherical and tangent spaces. In Fig. 2(c), spherical space and tangent space intersect at N, where $N = (0, ..., r) \in \mathbb{R}^n$ is both the north pole of the spherical space $\mathbb{S}_r^{n-1} = \{X \in \mathbb{R}^n : \|X\| = r\}$ and the origin of the tangent space T_N. Given a vector $\hat{m} = (m, 0)$ (*i.e.*, the blue solid line) in T_{N_1} starting at N, $\|\hat{m}\|$ on T_N is the same as the geodesic distance between N and X on $\mathbb{S}_r^{n-1}$ [48] (*i.e.*, the blue dotted line). Based on this fact, the spherical logarithmic mapping $\mathcal{LOG}_N : \mathbb{S}_r^{n_1-1} \to T_{N_1}$ (*i.e.*, the red solid line) can be expressed as:

$$\begin{aligned} \mathcal{LOG}_N(X) =& \frac{\alpha}{\sin\alpha}(X - N\cos\alpha), \quad \forall X \in \mathbb{S}_r^{n_1-1} \\ & \alpha = \arccos(N^T X/r^2), \end{aligned} \tag{2}$$

where X represents the feature in location (i, j).

Tangent Space Transform. The tangent space transform is designed to accomplish feature propagation between tangent spaces ($T_{N_1} \to T_{N_2}$), which is essentially a fully connected layer:

$$\mathcal{L}(\hat{m}, r) = (Wm + b, r), \quad \forall \hat{m} \in T_{N_1} \tag{3}$$

where $W \in \mathbb{R}^{(n_2-1)*(n_1-1)}$ represents a weight matrix and $b \in \mathbb{R}^{n_2-1}$ represents a bias.

Spherical Exponential Mapping. Similar to the spherical logarithmic mapping, the spherical exponential mapping $\mathcal{EXP}_N : T_{N_2} \to \mathbb{S}_r^{n_2-1}$ is also accomplished based on the fact that the geodesic distance from N to $(m, 0)$ and N to X are the same:

$$\begin{aligned} \mathcal{EXP}_N(\hat{m}) &= N\cos\beta + \hat{m}\frac{\sin\beta}{\beta}, \quad \forall \hat{m} \in T_{N_2} \\ \beta &= \|\hat{m}\|/r, \end{aligned} \tag{4}$$

For clarity, we provide mathematical proofs of spherical logarithmic mapping and spherical exponential mapping in Algorithm 1 and Algorithm 2.

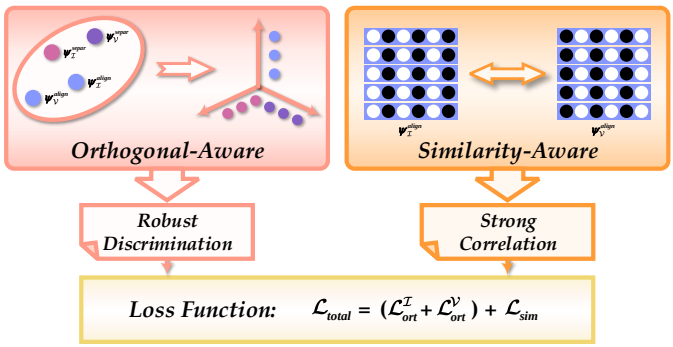

Fig. 4. The pipeline of the loss function in the shadow removal stage. The training loss consists of the orthogonality loss $\mathcal{L}_{ort}$ and the similarity loss $\mathcal{L}_{sim}$. The former separates independence by calculating the inner product between shared and private features. The latter builds a bridge to explore semantic similarities and differences between shared features.

B. Encoder–Decoder

The encoder employs cascaded Swin Transformer blocks [49] to simultaneously explore the global dependence of visible features $\mathcal{V}$ and infrared features $\mathcal{I}$. Swin Transformer is known for the long-range context exploitation capability of the multi-head self-attention mechanism, focusing on the capture of shared and private features in the channel dimension [45], as shown in Fig. 3. Taking visible features as an example, we assume that the former $\frac{dim}{2}$ channels are shared, representing cross-modal information. Obviously, the latter $\frac{dim}{2}$ channels are private, representing the texture of the visible object surface. We expect to leverage the infrared modality to provide illumination representations missing in the visible modality, thus domain-shared and domain-private features should be separated and aligned, respectively. To this end, we employ the logarithmic mapping $\mathcal{LOG}(\cdot)$ to project multi-modal features into the spherical space, accomplish the feature decomposition under the supervision of orthogonality and similarity losses, and reproduce the promising illumination through the exponential mapping $\mathcal{EXP}(\cdot)$:

$$\begin{aligned}
\psi_{\mathcal{M}}^{align} &= \mathcal{S}(\mathcal{M})[0:\frac{dim}{2}],\\
\psi_{\mathcal{M}}^{separ} &= \mathcal{S}(\mathcal{M})[\frac{dim}{2}:dim],\\
\hat{\psi}_{\mathcal{M}}^{align} &= \mathcal{T}(\psi_{\mathcal{M}}^{align}),\\
\hat{\psi}_{\mathcal{M}}^{separ} &= \mathcal{T}(\psi_{\mathcal{M}}^{separ}),\\
\hat{\psi}_{\mathcal{M}} &= \mathcal{C}(\hat{\psi}_{\mathcal{M}}^{align}, \hat{\psi}_{\mathcal{M}}^{separ}),
\end{aligned} \tag{5}$$

where $\mathcal{S}(\cdot)$ represents the Swin Transformer block and $\mathcal{C}(\cdot)$ represents the dimension concatenation operator.

For shadow removal, the full visible feature $\hat{\psi}_{\mathcal{V}}$ with rich edges and the shared infrared feature $\hat{\psi}_{\mathcal{I}}^{align}$ with robust illumination representations are beneficial. Therefore, we integrate $\hat{\psi}_{\mathcal{V}}$ and $\hat{\psi}_{\mathcal{I}}^{align}$ in the channel dimension to accomplish shadow contamination removal:

$$\mathcal{S}_f = \mathcal{S}(\mathcal{C}(\hat{\psi}_{\mathcal{V}}, \hat{\psi}_{\mathcal{I}}[0:\frac{dim}{2}])). \tag{6}$$

where $\mathcal{S}_f$ represents the shadow-free image.

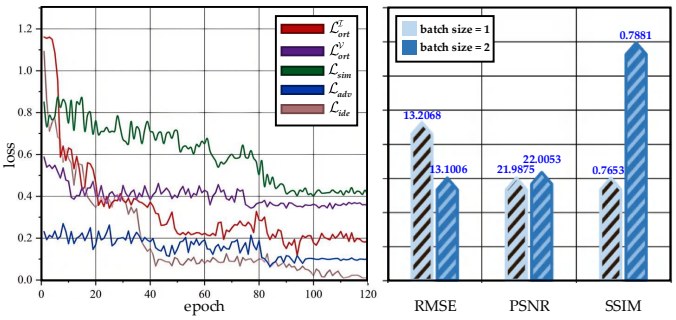

Fig. 5. Visualization towards epoch and batch size. Notably, a normalization process is applied to the RMSE, PSNR, and SSIM scores.

C. Loss Function

In order to achieve the decomposition of shared and private properties in a weakly supervised manner, we employ a linear combination of the orthogonality loss $\mathcal{L}_{ort}$, the similarity loss $\mathcal{L}_{sim}$, the adversarial loss $\mathcal{L}_{adv}$, and the identical loss $\mathcal{L}_{ide}$, which can be expressed as:

$$\begin{aligned}
\mathcal{L}_{total} &= \mathcal{L}_{ort}^{\mathcal{I}}(\psi_{\mathcal{I}}^{align}, \psi_{\mathcal{I}}^{separ}) + \mathcal{L}_{ort}^{\mathcal{V}}(\psi_{\mathcal{V}}^{align}, \psi_{\mathcal{V}}^{separ})\\
&+ \mathcal{L}_{sim}(\psi_{\mathcal{I}}^{align}, \psi_{\mathcal{V}}^{align}) + \mathcal{L}_{adv}(G, D) + \mathcal{L}_{ide}(G),
\end{aligned} \tag{7}$$

In Fig. 4, the orthogonality loss learns an orthogonal representation from the trend of change in angle, where two distinct properties are effectively separated, displaying ideally independent distributions. Specifically, we calculate the Gram matrices of shared and private features, and then straighten them to 1-D vectors. The orthogonality loss is defined as the inner product between vectors:

$$\begin{aligned}
\mathcal{L}_{ort}^{\mathcal{I}} &= \mathcal{F}(\mathcal{G}(\psi_{\mathcal{I}}^{align})) \cdot \mathcal{F}(\mathcal{G}(\psi_{\mathcal{I}}^{separ})),\\
\mathcal{L}_{ort}^{\mathcal{V}} &= \mathcal{F}(\mathcal{G}(\psi_{\mathcal{V}}^{align})) \cdot \mathcal{F}(\mathcal{G}(\psi_{\mathcal{V}}^{separ})),
\end{aligned} \tag{8}$$

where $\mathcal{F}$ represents the flattening operation and $\mathcal{G}$ represents the Gram matrix. The similarity loss promotes the consistency in illumination representations of multi-modal shared features:

$$\begin{aligned}
\text{SSIM} &= \frac{(2\mu_{\mathcal{I}}\mu_{\mathcal{V}} + c_1)(2\sigma_{\mathcal{IV}} + c_2)}{(\mu_{\mathcal{I}}^2 + \mu_{\mathcal{V}}^2 + c_1)(\sigma_{\mathcal{I}}^2 + \sigma_{\mathcal{V}}^2 + c_2)},\\
\mathcal{L}_{sim} &= 1 - \text{SSIM}(\psi_{\mathcal{I}}^{align}, \psi_{\mathcal{V}}^{align}),
\end{aligned} \tag{9}$$

where $\mu_{\mathcal{I}}$ and $\mu_{\mathcal{V}}$ represent the means of $\psi_{\mathcal{I}}^{align}$ and $\psi_{\mathcal{V}}^{align}$, respectively. $\sigma_{\mathcal{I}}^2$ and $\sigma_{\mathcal{V}}^2$ represent the corresponding variances. $\sigma_{\mathcal{IV}}$ represents the covariance of $\psi_{\mathcal{I}}^{align}$ and $\psi_{\mathcal{V}}^{align}$. The constants c_1 and c_2 are set to 0.01^2 and 0.03^2, avoiding numerical instability as the denominator approaches zero.

The adversarial loss promotes the shadow generation stage to learn the shadow distribution, while the identical loss encourages the contamination generation consistent with the cropped shadow S:

$$\begin{aligned}
\mathcal{L}_{adv} &= \mathbb{E}_{\mathcal{V},\mathcal{S}_f}[\log(1 - D(G(\mathcal{V})))] + \mathbb{E}_{\mathcal{V},\mathcal{S}_f}[\log D(\mathcal{S}_f)],\\
\mathcal{L}_{ide} &= \mathbb{E}_{\mathcal{V}\sim p(\mathcal{V})}[\|G(\mathcal{V}), S\|_1].
\end{aligned} \tag{10}$$

where $p(\cdot)$ represents the data distribution and $\|,\|_1$ represents the $\mathcal{L}_1$ loss.

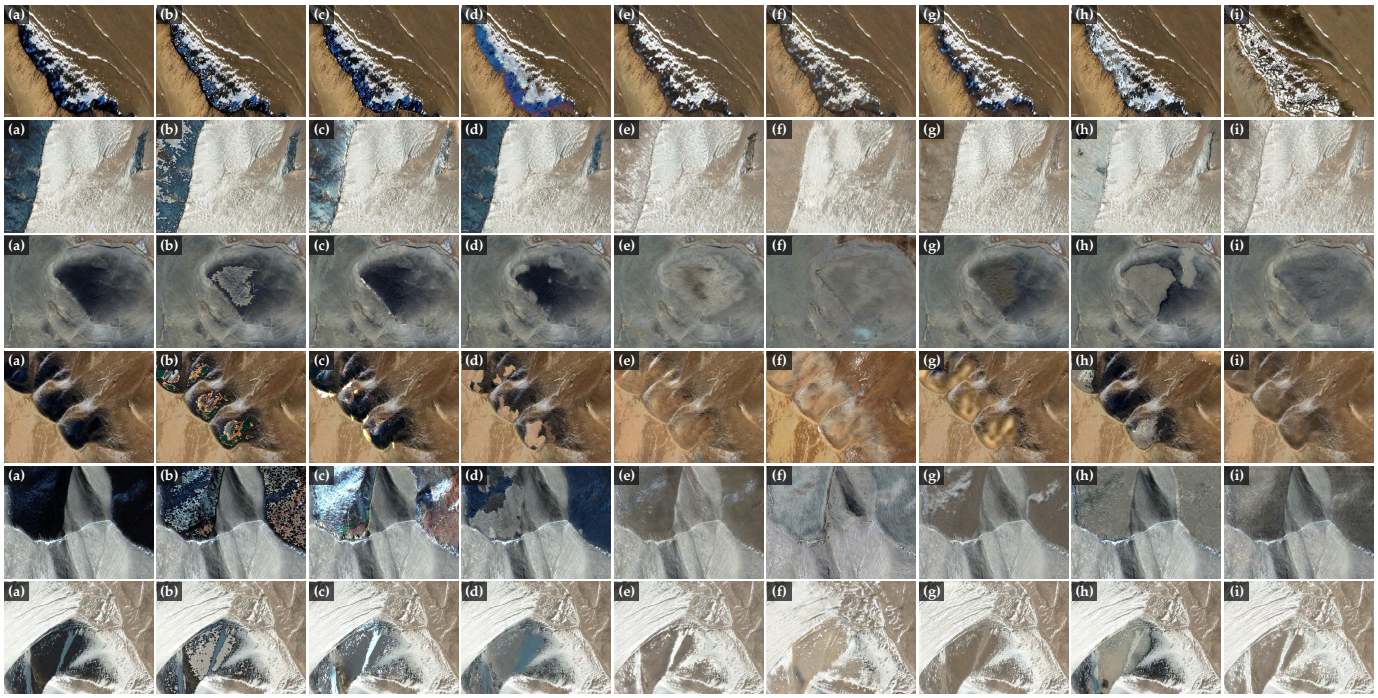

Fig. 6. Visual comparisons on shadow images sampled from **WSSR**. (a) input. (b) Silva [8]. (c) Gong [11]. (d) Guo [37]. (e) Mask-ShadowGAN [19]. (f) DC-ShadowNet [20]. (g) LG-ShadowNet [21]. (h) G2R-ShadowNet [23]. (i) S2-ShadowNet.

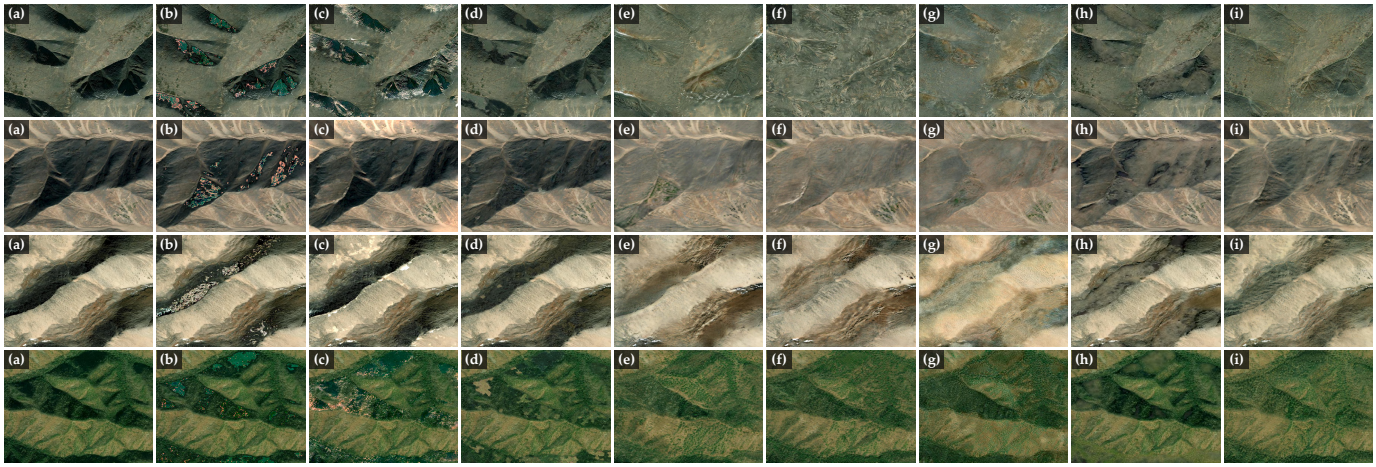

Fig. 7. Visual comparisons on shadow images sampled from **URSSR** [50]. (a) input. (b) Silva [8]. (c) Gong [11]. (d) Guo [37]. (e) Mask-ShadowGAN [19]. (f) DC-ShadowNet [20]. (g) LG-ShadowNet [21]. (h) G2R-ShadowNet [23]. (i) S2-ShadowNet.

IV. Experiment

A. Experimental Settings

Implementation Details. For the modal translation stage, we employ a pretrained U2Fusion [44] to provide infrared samples. U2Fusion defines the importance of visible features through richness measurement, and then generates visually realistic infrared modalities through weight assignment. For the shadow generation stage, we refer to [23]. Given a shadow mask representing the shadow region, we randomly crop the region of approximate area from the shadow-free portion. Subsequently, we employ a cycle constraint metwork [19] to add pseudo shadows to this cropped shadow-free region. At this point, the cropped shadow-free version and the pseudo shadow version form a supervised pair, thus providing training samples for subsequent shadow removal. For the shadow removal stage, relevant details have been provided in the methodology section. Notably, the shadow generation stage and shadow removal stage are trained from scratch. In summary,the implementation of S2-ShadowNet is done with PyTorch on an NVIDIA RTX 3090 GPU. The training epoch is set as 100. A batch-mode learning strategy with a batch size of 2 is employed. In Fig. 5, multiple loss curves become relatively smooth after 100 epochs, thus increasing the epoch neither improves the shadow removal performance. Besides, when the batch size is set to 2, the RMSE, PSNR, and SSIM scores are slightly higher than the performance when the batch size is set to 1. More importantly, setting the batch size to 2

Fig. 8. Visual comparisons on shadow images sampled from **UAV-SC** [51]. (a) input. (b) Silva [8]. (c) Gong [11]. (d) Guo [37]. (e) Mask-ShadowGAN [19]. (f) DC-ShadowNet [20]. (g) LG-ShadowNet [21]. (h) G2R-ShadowNet [23]. (i) DMTN [6]. (j) ST-CGAN [15]. (k) DHAN [39]. (l) ShadowFormer [40]. (m) TBRNet [42]. (n) S2-ShadowNet. (o) ground truth.

can save 30% of the training time cost. ADAM is applied for network optimization and the learning rate is fixed to $1e^{-4}$.

Benchmarks. We perform qualitative and quantitative comparisons on widely used WSSR, URSSR [50], and UAV-SC [51] benchmarks. **WSSR** consists of 1000 shadow images with 4K resolution from the Rocky, Appalachian, Caucasus, and Altai mountains, *etc*. For data augmentation, we randomly crop WSSR into 4000 samples. Notably, to enable weakly supervised and unsupervised learning, we also provide corresponding shadow masks [52] and non-consistent shadow-free samples. This further enhances the utility of WSSR. We randomly select 3900 shadow images and corresponding shadow masks for training, while the rest 100 shadow images for testing. **URSSR** contains unpaired 350 shadow and 230 shadow-free images captured from multiple national parks, such as Jirisan, Deogyusan, and Gayasan national parks. Similarly, we randomly select 300 shadow images while employing all shadow-free images to train S2-ShadowNet, and the rest 50 shadow images for testing. **UAV-SC** consists of 6954 shadow images and reference ground truths, where 6924 image pairs are used for training and the rest 30 image pairs are used for testing. UAV-SC is collected from Wuhan through two drone platforms. Besides, radiometric and geometric corrections are employed to support the consistency of image pairs. In summary, the training and testing sets cover diverse mountain, woodland, and urban scenarios, different degradation characteristics, and a broad range of image content.

Compared Methods. We compare S2-ShadowNet with a series of representative shadow removal methods, including **traditional-based methods** (Silva [8], Gong [11], and Guo [37]) and **deep learning-based methods** (Mask-ShadowGAN [19], DC-ShadowNet [20], LG-ShadowNet [21], G2R-ShadowNet [23], DMTN [6], ST-CGAN [15], DHAN [39], ShadowFormer [40], and TBRNet [42]).

Evaluation Metrics. For UAV-SC, we perform full-reference evaluations by computing the root mean square error (**RMSE**), the peak signal-to-noise ratio (**PSNR**), and the structural similarity (**SSIM**). A lower RMSE score suggests better performance, while the opposite is true for PSNR and SSIM scores. Notably, the full-reference evaluation is computed in

TABLE I
QUANTITATIVE COMPARISONS ON **UAV-SC** [51]. "↑" REPRESENTS THAT LARGER SCORES ARE BETTER, WHILE "↓" REPRESENTS THAT LOWER SCORES ARE BETTER. "−" REPRESENTS THE EXECUTION TIME IS NOT AVAILABLE. BEST AND SECOND-BEST SCORES ARE **HIGHLIGHTED** AND UNDERLINED. †, ‡, AND ¶ REPRESENT UNSUPERVISED, WEAKLY SUPERVISED, AND FULLY SUPERVISED METHODS, RESPECTIVELY.

Methods	RMSE(↓)			PSNR(↑)			SSIM(↑)			Time/s(↓)
	S.	N.S.	All	S.	N.S.	All	S.	N.S.	All	
Silva [8] (ISPRS'18)	32.0696	18.5089	21.3024	22.8902	18.1034	16.2621	0.8899	0.8035	0.6898	1.51
Gong [11] (BMVC'14)	26.0850	18.8162	20.3155	25.1448	18.1944	16.8290	0.9179	0.7926	0.6984	-
Guo [37] (TPAMI'13)	29.8381	17.9412	20.3919	23.5991	18.3971	16.6687	0.9042	0.8093	0.6945	0.37
Mask-ShadowGAN [19] (ICCV'19)†	19.7786	17.5036	17.9722	27.5729	20.4507	19.1357	0.9450	0.8142	0.7374	0.13
DC-ShadowNet [20] (ICCV'21)†	16.1495	13.6091	14.1324	29.2519	22.4207	21.0712	0.9543	0.8479	0.7852	0.09
LG-ShadowNet [21] (TIP'21)†	23.0353	16.6409	17.9581	25.4526	19.6434	18.0556	0.9352	0.8352	0.7629	0.10
G2R-ShadowNet [23] (CVPR'21+Mask)‡	22.0473	18.6220	19.3276	24.6951	17.9899	16.8325	0.9015	0.8189	0.7183	0.08
DMTN [6] (TMM'23+Mask)¶	11.5195	9.9282	10.2560	31.6098	24.1577	23.0287	0.9664	0.8696	0.8195	0.07
ST-CGAN [15] (CVPR'18+Mask)¶	15.0976	13.0098	13.4399	27.4801	21.5584	20.2338	0.9525	0.8176	0.7486	**0.06**
DHAN [39] (AAAI'20+Mask)¶	**9.1524**	**9.0602**	**9.1086**	**33.0654**	**25.1760**	**24.1399**	**0.9752**	**0.9010**	**0.8620**	0.07
ShadowFormer [40] (AAAI'23+Mask)¶	9.6703	9.2937	9.5927	32.7140	24.3702	23.4432	0.9711	0.8804	0.8350	0.09
TBRNet [42] (TNNLS'23+Mask)¶	11.4115	10.8361	10.9546	31.2298	23.6601	22.5798	0.9619	0.8313	0.7722	0.10
S2-ShadowNet	14.8059	12.9814	13.1006	30.7410	22.9772	22.0053	0.9570	0.8509	0.7881	**0.06**

TABLE II
QUANTITATIVE COMPARISONS ON **WSSR**. "↑" REPRESENTS THAT LARGER SCORES ARE BETTER. BEST AND SECOND-BEST SCORES ARE **HIGHLIGHTED** AND UNDERLINED. † AND ‡ REPRESENT UNSUPERVISED AND WEAKLY SUPERVISED METHODS.

Methods	VNM(↑)	Entropy(↑)	Time/s(↓)
Silva [8]	0.2461	6.5851	1.40
Gong [11]	0.2437	6.5510	-
Guo [37]	0.2459	6.6002	0.29
Mask-ShadowGAN [19] †	0.2581	6.7909	0.12
DC-ShadowNet [20] †	0.2572	6.7123	0.09
LG-ShadowNet [21] †	0.2682	6.8737	0.09
G2R-ShadowNet [23] ‡	0.2519	6.6778	0.08
S2-ShadowNet	**0.3062**	**6.9418**	**0.06**

TABLE III
QUANTITATIVE COMPARISONS ON **URSSR** [50]. "↑" REPRESENTS THAT LARGER SCORES ARE BETTER. BEST AND SECOND-BEST SCORES ARE **HIGHLIGHTED** AND UNDERLINED. † AND ‡ REPRESENT UNSUPERVISED AND WEAKLY SUPERVISED METHODS.

Methods	VNM(↑)	Entropy(↑)	Time/s(↓)
Silva [8]	0.2527	6.5902	1.34
Gong [11]	0.2540	6.6077	-
Guo [37]	0.2659	6.6216	0.33
Mask-ShadowGAN [19] †	0.2926	6.6470	0.12
DC-ShadowNet [20] †	0.2966	6.6160	0.09
LG-ShadowNet [21] †	0.3138	6.7738	0.10
G2R-ShadowNet [23] ‡	0.2868	6.5983	0.07
S2-ShadowNet	**0.3582**	**7.0356**	**0.06**

the shadow region (S.), non-shadow region (N.S.), and whole image (All), respectively. For WSSR and URSSR without ground truths, we employ visual neuron matrix (**VNM**) [53] and **Entropy** to perform no-reference evaluations. VNM extracts visual features from shadow removal images by simulating visual neurons in the cerebral cortex. Then, the neural network is used to associate visual features with corresponding quality scores. Entropy serves as a statistical tool for visual features, which typically reflects the information richness. A higher VNM or Entropy score indicates a more attractive visual perception.

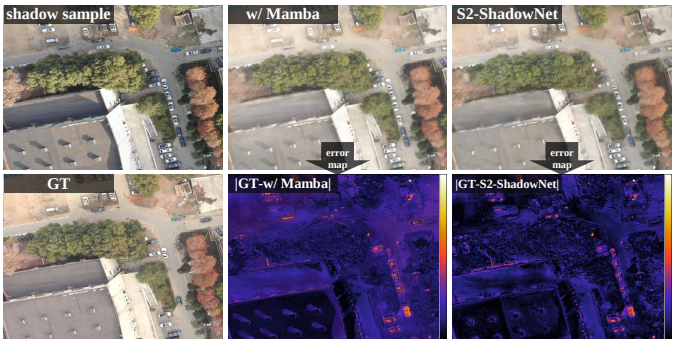

Fig. 9. Ablation study towards the encoder–decoder strategy. Error maps suggest that Swin Transformer is more effective as a encoder–decoder backbone for small-scale shadow casts.

B. Visual Comparisons

We first show the comparisons on shadow images with complex cast-surface geometric shapes and boundaries sampled from WSSR in Fig. 6. All traditional-based methods fail to cope with shadow effects, *e.g.*, Silva [8] introduces color artifacts, Gong [11] changes original tones, and Guo [37] produces local degradation. Although Mask-ShadowGAN [19], DC-ShadowNet [20], and LG-ShadowNet [21] remove shadow traces to some extent, the consistency and coordination of shadow and shadow-free regions fail to be repaired. In addition, G2R-ShadowNet [23] introduces distorted illumination compensation, which seriously destroys the structure and pattern of the remote sensing imagery. In contrast, our method removes shadows while maintaining color consistency, which is credited to the fact that the infrared modality provides crucial illumination prior for the visible modality.

We then show the comparisons on challenging shadow images sampled from URSSR [50] in Fig. 7. Silva [8] introduces reddish and greenish color deviations. Non-homogeneous shadow distribution challenges Gong [11] and Guo [37]. This is because 1) accurate annotations are demanding for users, 2) inaccurate shadow detection exacerbates unsatisfactory performance. G2R-ShadowNet [23] fails to handle continuously varying shadow intensities, thus the robustness is not convinc-

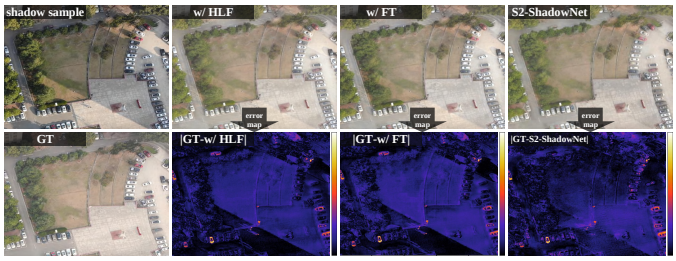

Fig. 10. Ablation study towards the feature decomposition. Error maps suggest that channel dimension decomposition is simple and effective for large-scale shadow traces.

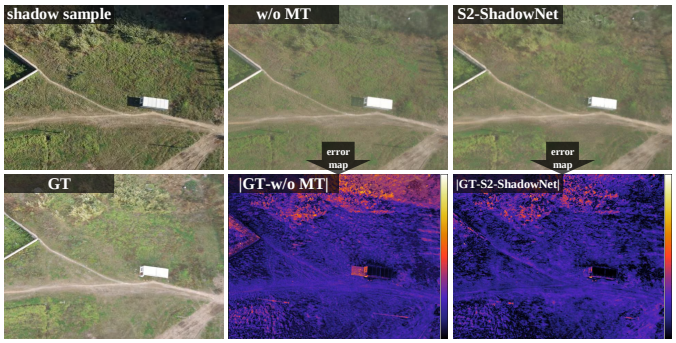

Fig. 11. Ablation study towards the modal translation. Error maps suggest that the introduction of infrared modality contributes to remove shadow traces.

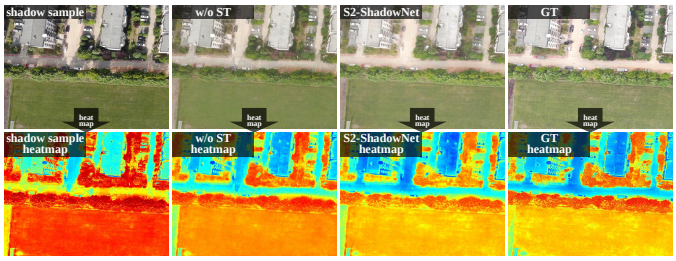

Fig. 12. Ablation study towards the spherical transform. Heatmaps suggest that the aggregation of multiple modalities in spherical space achieves a double win across illumination compensation and detail restoration.

ing. Mask-ShadowGAN [19], DC-ShadowNet [20], and LG-ShadowNet [21] are attractive for illumination recovery, but the texture details of the mountains are inevitably hidden. In contrast, our method not only relights the dark pixels, but also maintains realistic and rich details.

We also show the comparisons on shadow images with illumination variations and shadow overlaps sampled from UAV-SC [51] in Fig. 8. Cast shadow is tricky for Silva [8], Gong [11], and Guo [37] because tiny shadow remnants are difficult to localize implicitly. Competing unsupervised [19]–[21] and weakly supervised [23] methods either render shadow remnants or introduce artifacts. In addition, some fully supervised methods fail to maximally preserve detail information, such as DMTN [6] and ST-CGAN [15]. In contrast, our method, DHAN [39], ShadowFormer [40], and TBRNet [42] are closer to ground truths.

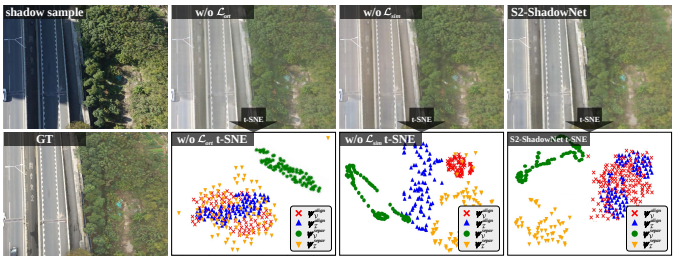

Fig. 13. Ablation study towards the orthogonality loss $\mathcal{L}_{ort}$ and the similarity loss $\mathcal{L}_{sim}$. t-SNE suggests that the collaboration between orthogonality loss and similarity loss gracefully accomplishes cross-modal feature decomposition.

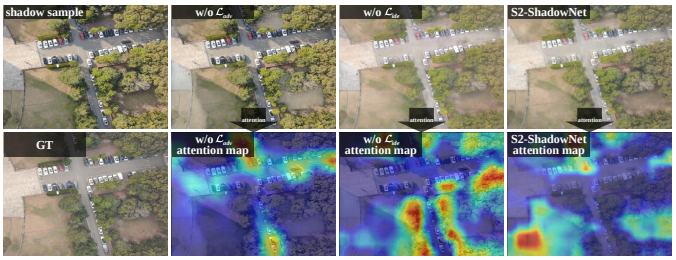

Fig. 14. Ablation study towards the adversarial loss $\mathcal{L}_{adv}$ and the identical loss $\mathcal{L}_{ide}$. Feature visualizations suggest that the interaction between adversarial loss and identical loss focuses more attention on shadow remnants.

TABLE IV
QUANTITATIVE SCORES OF THE ABLATION STUDY IN TERMS OF RMSE, PSNR, AND SSIM SCORES. "↑" REPRESENTS THAT LARGER SCORES ARE BETTER, WHILE "↓" REPRESENTS THAT LOWER SCORES ARE BETTER. THE BEST SCORE IS **HIGHLIGHTED**.

Baselines	UAV-SC		
	RMSE(↓)	PSNR(↑)	SSIM(↑)
w/ Mamba	13.8626	21.7103	0.7810
w/ HLF	15.3039	21.0039	0.7522
w/ FT	15.1198	21.2891	0.7739
w/o MT	19.7614	16.8275	0.7212
w/o ST	16.5299	20.8440	0.7483
w/o $\mathcal{L}_{ort}$	15.3131	20.9069	0.7608
w/o $\mathcal{L}_{sim}$	14.8689	21.0520	0.7768
w/o $\mathcal{L}_{adv}$	19.9431	16.7428	0.7341
w/o $\mathcal{L}_{ide}$	17.3769	19.8943	0.7391
full model	**13.1006**	**22.0053**	**0.7881**

C. Quantitative Comparisons

For fair quantitative comparisons, we employ the source code provided by the authors, then retrain the compared methods using the consistent training set and achieve the best quantitative scores. We first report the average RMSE, PSNR, and SSIM scores of the developed and compared methods in Table I. Our method outperforms traditional-based, unsupervised, and weakly supervised methods on the UAV-SC benchmark. Admittedly, the fully supervised method has overwhelming superiority over the weakly supervised method in full-reference evaluations [54], [55]. Our method is comparable to some fully supervised methods [15]. Coupled with the rigorous and expensive data acquisition of the fully supervised paradigm, our method is encouraging.

We then report the average VNM and Entropy scores on WSSR and URSSR benchmarks in Tables II and III. Compared with the top-performing LG-ShadowNet [21], our method

achieves the percentage gain of 14.2%/0.9% and 14.1%/3.9% in terms of VNM/Entropy on WSSR and URSSR, respectively. Such quantitative scores demonstrate the superiority of our method for shadow removal. More importantly, our method presents the best execution time on the UAV-SC [51], WSSR, and URSSR [50] benchmarks, which implies that S2-ShadowNet is convincingly efficient. Notably, we do not provide the execution time of Gong [11]. This is because Gong [11] asks the user to draw shadow contours at runtime. This leads to lengthy execution times and is highly dependent on experience and proficiency.

D. Ablation Studies

We conduct extensive ablation studies to analyze the core components of S2-ShadowNet, including the modal translation and the spherical transform. In addition, we analyze the linear combination of the orthogonality loss $\mathcal{L}_{ort}$, the similarity loss $\mathcal{L}_{sim}$, the adversarial loss $\mathcal{L}_{adv}$, and the identical loss $\mathcal{L}_{ide}$. More specifically,

- w/ Mamba refers to S2-ShadowNet employs Mamba [61] instead of Swin Transformer as the encoder–decoder strategy.
- w/ HLF and w/ FT refer to S2-ShadowNet employs high-low frequency decomposition [62] and fourier transform [63] instead of channel dimension decomposition, respectively. Notably, the quality degradation is stored in the high-frequency component and the amplitude component, respectively.
- w/o MT refers to S2-ShadowNet without the modal translation, similar to [23], employing shadow and random shadow-free regions to accomplish weakly supervised learning.
- w/o ST refers to S2-ShadowNet without the spherical transform, instead employing feature-wise concatenation to integrate visible and infrared representations.
- w/o $\mathcal{L}_{ort}$ means that S2-ShadowNet is trained without the constraint of the orthogonality loss.
- w/o $\mathcal{L}_{sim}$ means that S2-ShadowNet is trained without the constraint of the similarity loss.
- w/o $\mathcal{L}_{adv}$ means that S2-ShadowNet is trained without the constraint of the adversarial loss.
- w/o $\mathcal{L}_{ide}$ means that S2-ShadowNet is trained without the constraint of the identical loss.

The quantitative scores of the ablated models on UAV-SC are reported in Table IV. The effectiveness of encoder–decoder and feature decomposition are shown in Figs. 9 and 10. In addition, the contributions of modal translation, the effectiveness of spherical transform, and the effects of loss function are depicted in Figs. 11, 12, 13, and 14, respectively. The conclusions drawn from the ablation study are listed as follows.

- As shown in Table IV, the full model achieves the best RMSE, PSNR, and SSIM scores, which implies that the introduction of infrared modality and the design of spherical space transform are convincing.
- In Fig. 9, the ablated model w/ Mamba fails to remove shadow casts on the roof. This is because Mamba flattens the spatial data, which disrupts local dependencies. In contrast, Swin Transformer performs self-attention in local windows, which ensures that S2-ShadowNet can flexibly capture small-scale shadow casts. In Fig. 10, the ablated models w/ HLF and w/ FT still retain shadow remnants. This is because the high-low frequency decomposition treats shadows as noise masks. However, brute force noise separation struggles to accommodate mixing patterns with diverse noises and image contents [62]. Besides, the core insight of the Fourier transform is to explore the positive correlation between amplitude and illumination (ground truth), and providing only an infrared complement hardly fulfills the potential of the Fourier transform. In contrast, the division between shared and private features is channel-wise [48], and the channel dimension decomposition fits seamlessly with such division. More importantly, the effectiveness of channel dimension decomposition has been demonstrated in [45].
- In Fig. 11, the ablated model w/o MT produces undesired shadow traces, which are obvious in error map visualizations. In contrast, the full model removes shadow boundaries through the close cooperation of visible and infrared information.
- Arbitrary and diverse cast shadows challenge the ablated model w/o ST, as shown in Fig. 12. The heatmap indicates that aimlessly integrating multi-modal information fails to recover promising illumination. In contrast, the heatmap of the full model is closer to the corresponding ground truth in terms of detail and texture. Therefore, employing spherical space to perform domain aggregation is imperative.
- As shown in Fig. 13, the ablated models w/o $\mathcal{L}_{ort}$ and w/o $\mathcal{L}_{sim}$ either fail to separate private and shared features or fail to align shared features. In contrast, the full model integrates multi-modal shared features while pushing them away from multi-modal private features for attractive cross-modal modeling.
- In Fig. 14, we observe obvious shadow effects without employing the adversarial loss $\mathcal{L}_{adv}$. This is because the lack of $\mathcal{L}_{adv}$ fails to provide shadow generation that supports weakly supervised learning. Besides, the ablated model w/o $\mathcal{L}_{ide}$ fails to focus attention on shadow regions, leading to shadow remnants. The full model allocates more shadow-aware attention to quality-degraded regions, thus achieving visually pleasing quality.

V. Conclusion

In this work, we propose a weakly supervised shadow removal network that aims to free users from tedious data pair acquisition. The core insight of S2-ShadowNet is to employ the infrared modality to provide an additional illumination prior for the visible modality, thus enjoying an intimate cooperation between multiple modalities. To this end, we introduce a spherical space to bridge the inter-domain differences. The main reason for choosing the spherical space is that the distances between spherical features can be regularized, which

helps to select an appropriate distance measure for feature alignment and separation. Under the supervision of well-designed orthogonality loss and similarity loss, the private features between infrared and visible modalities complete the independence separation through inner product while the shared features accomplish the semantic level alignment. Such a manner achieves a deep cross-modal fusion. Extensive experiments have demonstrated the superiority and efficiency of S2-ShadowNet.

In future research, we plan to upgrade S2-ShadowNet to a multi-task pipeline, which applies it to downstream tasks such as urban road extraction and vegetation monitoring. Such a manner helps to broaden the utility of our methodology. In addition, employing a game engine to produce fully supervised shadow removal benchmarks is also a promising direction we are exploring.

REFERENCES

[1] Q. Meng, S. Zhang, Z. Li, C. Wang, W. Zhang, and Q. Huang, "Automatic shadow generation via exposure fusion," *IEEE Trans. Multimedia*, vol. 25, pp. 9044–9056, Feb. 2023.

[2] J. Shen, C. Zhang, Y. Yuan, and Q. Wang, "Enhancing prospective consistency for semisupervised object detection in remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–12, Aug. 2023.

[3] W. Jing, Y. Yuan, and Q. Wang, "Dual-field-of-view context aggregation and boundary perception for airport runway extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–12, May. 2023.

[4] Y. Jia, J. Gao, W. Huang, Y. Yuan, and Q. Wang, "Holistic mutual representation enhancement for few-shot remote sensing segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–13, Oct. 2023.

[5] Y. Jia, J. Gao, W. Huang, Y. Yuan, and Q. Wang, "Exploring hard samples in multiview for few-shot remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–14, Jul. 2023.

[6] J. Liu, Q. Wang, H. Fan, W. Li, L. Qu, and Y. Tang, "A decoupled multi-task network for shadow removal," *IEEE Trans. Multimedia*, vol. 25, pp. 9449–9463, Mar. 2023.

[7] G. D. Finlayson, S. D. Hordley, C. Lu, and M. S. Drew, "On the removal of shadows from images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 59–68, Jan. 2006.

[8] G. F. Silva, G. B. Carneiro, R. Doth, L. A. Amaral, and D. F. G. de Azevedo, "Near real-time shadow detection and removal in aerial motion imagery application," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 104–121, Jun. 2018.

[9] G. D. Finlayson, M. S. Drew, and C. Lu, "Entropy minimization for shadow removal," *Int. J. Comput. Vis.*, vol. 85, pp. 35–57, May. 2009.

[10] C. R. Jung, "Efficient background subtraction and shadow removal for monochromatic video sequences," *IEEE Trans. Multimedia*, vol. 11, no. 3, pp. 571–577, Apr. 2009.

[11] H. Gong and D. Cosker, "Interactive shadow removal and ground truth for variable scene categories," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Sep. 2014, pp. 1–11.

[12] K. Niu, Y. Liu, E. Wu, and G. Xing, "A boundary-aware network for shadow removal," *IEEE Trans. Multimedia*, vol. 25, pp. 6782–6793, Oct. 2023.

[13] Y. Liu *et al.*, "Structure-informed shadow removal networks," *IEEE Trans. Image Process.*, vol. 32, pp. 5823–5836, Oct. 2023.

[14] Y. Xu, M. Lin, H. Yang, F. Chao, and R. Ji, "Shadow-aware dynamic convolution for shadow removal," *Pattern Recognit.*, vol. 146, pp. 109969, Feb. 2024.

[15] J. Wang, X. Li, and J. Yang, "Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 1788–1797.

[16] H. Le and D. Samaras, "Shadow removal via shadow image decomposition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8577–8586.

[17] H. Le and D. Samaras, "Physics-based shadow image decomposition for shadow removal," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9088–9101, Dec. 2022.

[18] L. Fu *et al.*, "Auto-exposure fusion for single-image shadow removal," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10566–10575.

[19] X. Hu, Y. Jiang, C.-W. Fu, and P.-A. Heng, "Mask-ShadowGAN: Learning to remove shadows from unpaired data," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2472–2481.

[20] Y. Jin, A. Sharma, and R. T. Tan, "DC-ShadowNet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5007–5016.

[21] Z. Liu, H. Yin, Y. Mi, M. Pu, and S. Wang, "Shadow removal by a lightness-guided network with training on unpaired data," *IEEE Trans. Image Process.*, vol. 31, pp. 1853–1865, Jan. 2021.

[22] H. Le and D. Samaras, "From shadow segmentation to shadow removal," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Nov. 2020, pp. 264–281.

[23] Z. Liu, H. Yin, X. Wu, Z. Wu, Y. Mi, and S. Wang, "From shadow generation to shadow removal," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4925–4934.

[24] D. Lee, M. Jeon, Y. Cho, and A. Kim, "Edge-guided multi-domain RGB-to-TIR image translation for training vision tasks with challenging labels," in *Proc. IEEE Conf. Comput. Robot. Autom. (ICRA)*, May. 2023, pp. 8291–8298.

[25] L. Zhang, A. Gonzalez-Garcia, J. van de Weijer, M. Danelljan, and F. S. Khan, "Synthetic data generation for end-to-end thermal infrared tracking," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1837–1850, Apr. 2019.

[26] C. Li, W. Xia, Y. Yan, B. Luo, and J. Tang, "Segmenting objects in day and night: Edge-conditioned CNN for thermal image semantic segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 7, pp. 3069–3082, Jul. 2021.

[27] C. Devaguptapu, N. Akolekar, M. M. Sharma, and V. N. Balasubramanian, "Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1029–1038.

[28] V. V. Kniaz, V. A. Knyaz, J. Hladuvka, W. G. Kropatsch, and V. Mizginov, "ThermalGAN: Multimodal color-to-thermal image translation for person reidentification in multispectral dataset," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*, Nov. 2018, pp. 606–624.

[29] L. Gan, C. Lee, and S.-J. Chung, "Unsupervised RGB-to-Thermal domain adaptation via multi-domain attention network," in *Proc. IEEE Conf. Comput. Robot. Autom. (ICRA)*, May. 2023, pp. 6014–6020.

[30] B. Cao, Y. Sun, P. Zhu, and Q. Hu, "Multi-modal gated mixture of local-to-global experts for dynamic image fusion," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 23498–23507.

[31] Y. Zhang and H. Wang, "Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2153–2162.

[32] O. Mazhar, R. Babuska, and J. Kober, "GEM: Glare or gloom, I can still see you – end-to-end multi-modal object detection," *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 6321–6328, Oct. 2021.

[33] Q. Xie, T.-Y. Cheng, Z. Dai, V. Tran, N. Trigoni, and A. Markham, "Illumination-aware hallucination-based domain adaptation for thermal pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 1, pp. 315–326, Jan. 2024.

[34] J. Lee, S. Kim, S. Kim, and K. Sohn, "Multi-modal recurrent attention networks for facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 6977–6991, May. 2020.

[35] L. Jian, R. Rayhana, L. Ma, S. Wu, Z. Liu, and H. Jiang, "Infrared and visible image fusion based on deep decomposition network and saliency analysis," *IEEE Trans. Multimedia*, vol. 24, pp. 3314–3326, Jul. 2021.

[36] E. Arbel and H. Hel-Or, "Shadow removal using intensity surfaces and texture anchor points," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 6, pp. 1202–1216, Jun. 2011.

[37] R. Guo, Q. Dai, and D. Hoiem, "Paired regions for shadow detection and removal," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2956–2967, Dec. 2013.

[38] L. Zhang, Q. Zhang, and C. Xiao, "Shadow remover: Image shadow removal based on illumination recovering optimization," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4623–4636, Nov. 2015.

[39] X. Cun, C.-M. Pun, and C. Shi, "Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting GAN," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 10680–10687.

[40] L. Guo, S. Huang, D. Liu, C. Hao, and B. Wen, "ShadowFormer: Global context helps shadow removal," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 1, 2023, pp. 710–718.

[41] J. Wan, H. Yin, Z. Wu, X. Wu, Y. Liu, and S. Wang, "Style-guided shadow removal," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Nov. 2022, pp. 361–378.
[42] J. Liu, Q. Wang, H. Fan, J. Tian, and Y. Tang, "A shadow imaging bilinear model and three-branch residual network for shadow removal," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 11, pp. 15857–15871, Nov. 2024.
[43] Y. Zhu, H. Jie, X. Fu, F. Zhao, Q. Sun, and Z.-J. Zha, "Bijective mapping network for shadow removal," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5617–5626.
[44] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–518, Jan. 2022.
[45] Z. Zhao *et al.*, "Spherical space feature decomposition for guided depth map super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 12513–12524.
[46] X. Gu, J. Sun, and Z. Xu, "Spherical space domain adaptation with robust pseudo-label loss," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9098–9107.
[47] X. Gu, J. Sun, and Z. Xu, "Unsupervised and semi-supervised robust spherical space domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 3, pp. 1757–1774, Mar. 2024.
[48] Y. Shang *et al.*, "HDSS-Net: A novel hierarchically designed network with spherical space classifier for ship recognition in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–20, Nov. 2023.
[49] Z. Liu *et al.*, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
[50] Q. Wang, K. Chi, W. Jing, and Y. Yuan, "Recreating brightness from remote sensing shadow appearance," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–11, May. 2024.
[51] S. Luo, H. Li, Y. Li, C. Shao, H. Shen, and L. Zhang, "An evolutionary shadow correction network and a benchmark UAV dataset for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–14, Jul. 2023.
[52] L. Zhu *et al.*, "Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Nov. 2018, pp. 122–137.
[53] H.-W. Chang, X.-D. Bi, and C. Kai, "Blind image quality assessment by visual neuron matrix," *IEEE Signal Process. Lett.*, vol. 28, pp. 1803–1807, Aug. 2021.
[54] Q. Li, Y. Yuan, X. Jia, and Q. Wang, "Dual-stage approach toward hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 31, pp. 7252–7263, Nov. 2022.
[55] Q. Li, M. Gong, Y. Yuan, and Q. Wang, "RGB-induced feature modulation network for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–11, May. 2023.
[56] T. Hu, Q. Yan, Y. Qi, and Y. Zhang, "Generating content for HDR deghosting from frequency view," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 25732–25741.
[57] A. A. Khan *et al.*, "ORAN-B5G: A next-generation open radio access network architecture with machine learning for beyond 5G in industrial 5.0," *IEEE Trans. Green Commun. Netw.*, vol. 8, no. 3, pp. 1026–1036, Sep. 2024.
[58] A. A. Khan *et al.*, "Secure remote sensing data with blockchain distributed ledger technology: A solution for smart cities," *IEEE Access*, vol. 12, pp. 69383–69396, May. 2024.
[59] F. Mehmood, A. A. Khan, H. Wang, S. Karim, U. Khalid, and F. Zhao, "BLPCA-ledger: A lightweight plenum consensus protocols for consortium blockchain based on the hyperledger indy," *Comput. Stand. Interfaces*, vol. 91, Jan. 2025, Art. no. 103876.
[60] A. A. Khan, S. Dhabi, J. Yang, W. Alhakami, S. Bourouis, and P. L. Yee, "B-LPoET: A middleware lightweight proof-of-elapsed time (PoET) for efficient distributed transaction execution and security on blockchain using multithreading technology," *Comput. Electr. Eng.*, vol. 118, Aug. 2024, Art. no. 109343.
[61] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2023, *arXiv: 2312.00752*.
[62] C. Wang, Z. Zheng, R. Quan, Y. Sun, and Y. Yang, "Context-aware pretraining for efficient blind image decomposition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18186–18195.
[63] K. Chi, J. Li, W. Jing, Q. Li, and Q. Wang, "Neural implicit fourier transform for remote sensing shadow removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–10, Jun. 2024.

Kaichen Chi received the B.E. degree in electronic and information engineering and the M.E. degree in communication and information system from Liaoning Technical University, Huludao, China, in 2019 and 2022 respectively. He is currently working toward the Ph.D. degree in the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include image processing and deep learning.

Wei Jing received the B.M. degree in e-commerce and the M.S. degree in computer software and theory from Shandong University of Science and Technology, Qingdao, China, in 2019 and 2022 respectively. He is currently working toward the Ph.D. degree in the National Elite Institute of Engineering and the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include remote sensing image processing and deep learning.

Junjie Li received the B.E. degree in software engineering from Zhengzhou University, Zhengzhou, China, in 2024. He is currently working toward the M.S. degree in the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, pattern recognition and remote sensing.

Qiang Li (Member, IEEE) is currently with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University. His research interests include remote sensing image processing, particularly for image quality enhancement, object/change detection.

Qi Wang (Senior Member, IEEE) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, pattern recognition and remote sensing. For more information, visit the link (https://crabwq.github.io/).