Hidden Path Selection Network for Semantic Segmentation of Remote Sensing Images

Kunping Yang, Xin-Yi Tong, Gui-Song Xia, Senior Member, IEEE,
Weiming Shen, Liangpei Zhang, Fellow, IEEE

Abstract—Targeting at depicting land covers with pixel-wise semantic categories, semantic segmentation in remote sensing images needs to portray diverse distributions over vast geographical locations, which is difficult to be achieved by the homogeneous pixel-wise forward paths in the architectures of existing deep models. Although several algorithms have been designed to select pixel-wise adaptive forward paths for natural image analysis, it still lacks theoretical supports on how to obtain optimal selections. In this paper, we provide mathematical analyses in terms of the parameter optimization, which guides us to design a method called Hidden Path Selection Network (HPS-Net). With the help of hidden variables derived from an extra mini-branch, HPS-Net is able to tackle the inherent problem about inaccessible global optimums by adjusting the direct relationships between feature maps and pixel-wise path selections in existing algorithms, which we call hidden path selection. For the better training and evaluation, we further refine and expand the 5-class Gaofen Image Dataset (GID-5) to a new one with 15 land-cover categories, i.e., GID-15. The experimental results on both GID-5 and GID-15 demonstrate that the proposed modules can stably improve the performance of different deep structures, which validates the proposed mathematical analyses.

Index Terms—Remote sensing image, semantic segmentation, hidden path selection, benchmark dataset.

I. INTRODUCTION

SEMANTIC segmentation in remote sensing images [1], [2], [3], [4], which targets at identifying land covers with pixel-wise semantic categories, is of considerable significance among many land use management related tasks [5], [6], [7]. Due to the complexity of real-world environments, tackling issues such as the diversity of land-cover distributions over different geographical locations is an important topic among existing investigations that are related to remote sensing image interpretation [8], [9], [10], [11], [12], [13], [14]. As a consequence, studying semantic segmentation methods to depict diverse complicated land-cover distributions simultaneously has drawn increasing attentions in the remote sensing community [15], [16], [17].

For implementing semantic segmentation, the methods to transfer the spectral signals into digital features first came into researchers’ views, and classical classifiers in machine learning [18], [19] could be applied to predict pixel-wise land-cover categories [17], [20], [21], where elaborate computation procedures could be exploited to obtain features corresponding to diverse land-cover distributions. Nevertheless, with pre-defined man-made calculations, hand-crafted features are hard to represent numberless different land-cover distributions in real-world environments, which has encouraged following researches to focus on improving the flexibility and representation capabilities of features through learnable deep structures [3], [22], [23].

However, although optimal parameters could be adapted to land-cover distributions existing in training data samples, the deep structures are often static after being learned in the training processes, which could lose the adaptability to complicated real-world environments during the inference procedures [24]. In order to address this issue, many researchers have designed various elaborate structures to capture complex contextual information [25], [26], [27] not only for remote sensing images but also for natural ones.

Subsequently, neural architecture search (NAS) [28] is proposed to develop models with automatically acquired struc-
tuteres to adapt to diverse data distributions. Furthermore, some researchers have proposed structures, called dynamic routing, to select forward paths adaptive to each data sample through designed activating factors [24] instead of adapting to data distributions of the whole dataset as in [28]. However, with all pixels sharing the same activating factors, such methods might ignore the diversity over different image areas, the effect of which is remarkable in remote sensing images due to the vast covered geographical areas with diverse land-cover distributions. Consequently, the Gated Path Selection Network (GPSNet) [29] has been proposed to select pixel-wise adaptive forward paths, simply shown at the top of Fig.1, through designed soft masks instead of same activating factors shared by all pixels, as in [24]. However, considering all items in GPSNet as the variables, the direct functional relationships between soft masks (executing pixel-wise path selections through multiplication) and corresponding feature maps would make the parameter optimization constrained on a narrow high-dimension manifold instead of the full space, which might exclude the global optimums as illustrated in Fig.1. In other words, the direct relationships between feature maps and path selections could make it difficult to obtain the optimal pixel-wise forward path selections that are adaptive to different land-cover distributions. In summary, how to depict the land-cover distributions in complicated real-world environments for different pixels during the inference procedure has not been solved well by existing algorithms, which is of great interest to investigate.

In this paper, we portray the complicated pixel-wise land-cover distributions by imitating the hidden markov chain. As illustrated in Fig.2, we propose a method, called Hidden Path Selection Network (HPS-Net), with independent parallel structures (i.e. a main-branch for semantic segmentation and a mini-branch for supporting) by exploiting a sequence of special modules called Hidden Path Module (HP-Module). Considering all items in the main-branch as variables, we notice that the high-dimension manifold where we execute the parameter optimization could be adjusted through utilizing hidden variables highly independent with main-branch. Thus, we utilize the HP-Module to obtain a sequence of soft masks by leveraging hidden variables from the mini-branch to select pixel-wise adaptive forward paths in the main branch through multiplication, where the adjusted high-dimension manifold would have higher possibility to contain the global optimums.

Further inspired by the mathematical analyses in terms of parameter optimization, we cut off the gradient flows from soft masks to the main deep structures during feature integration in the HP-Module to improve the training efficiency, while the gradient flows are retained intactly during the multiplication. Then, according to the Taylor expansion with respect to soft masks in HPS-Net, whose values are defined within a small range, the loss function can approximate a convex function to some extent if global optimums are contained in the aforementioned high-dimension manifold. As a consequence, the usage of nearly freely valuing hidden variables would make soft masks adjusted freely and thus the convexity could help to obtain optimal soft masks (namely optimal pixel-wise path selections). As we shall see, with the proposed modules applied on existing deep structures, HPS-Net can better portray the pixel-wise land-cover distributions during the inference procedure.

Another important aspect related to semantic segmentation in remote sensing images is the accessibility of well-annotated dataset. Although images in many existing datasets, e.g.[30], [31], [32], [33], contain increasingly abundant land covers to compose different distributions, the included scenes can not perfectly reflect the complex real-world environments due to the limited data sample quantities, which would constrain the generalization abilities of trained deep models. Recently, a few large-scale datasets, e.g. [34], [35], have been built, but the involved land-cover categories are often insufficient yet and hardly meet the practical demands well. In this paper, we expand the land-cover classification dataset in [3] by refining the 5 involved land-cover categories in the whole GID-5 [34] into 15 categories, which can better serve the practical demands.

Our main contributions in this paper are threefold.

- We propose a new method, i.e. HPS-Net, to imitate hidden markov chain, which depicts land-cover distributions and selects adaptive forward paths for each pixel through an extra mini-branch highly independent with arbitrary main deep structures.
- We provide mathematical analyses on the parameter optimization with taking all items in deep structures as variables to explain the inaccessibility of global optimums, guided by which we design detailed structures of the HPS-Net and propose the HP-Module to improve the possibility to obtain the global optimums.
- We expand the large-scale remote sensing semantic segmentation dataset, i.e. GID-5, into 15 land-cover categories and propose the new GID-15 dataset, which provides a challenging large-scale benchmark platform with abundant land-cover categories and better satisfies the real-world scenarios.

II. RELATED WORK

A. Semantic Segmentation

Early researches in remote sensing semantic segmentation often leveraged and developed the digital features, among which some works [17], [20], [21] rely on the physical properties of land-covers and some other works [36], [37] capture the local spatial information. Subsequently, many deep structures [38], [39], [40], [41], whose fully connected layers are removed, have been borrowed in semantic segmentation models not only for remote sensing images but also for natural images [42], [43], [44]. Numerous learnable parameters in deep structures can help obtain features with stronger representation capabilities. Especially, researches on depicting different object distributions, such as the existence of multi-scale objects, have been hot spots for a long time [25], [26], [27]. Meanwhile, the community were inspired to research on neural architecture search (NAS) [28] for developing semantic segmentation models with automatically acquired structures.
Recently, some works have been proposed to focus on developing dynamic structures instead of static architectures (even automatically acquired). Specifically, activating factors obtained by designed soft conditional gates in [24] dynamically emphasize adaptive forward paths during the inference procedures. However, land-cover distributions often vary significantly within one remote sensing image due to the wide range of covered areas. Thus, the diversity of land-cover distributions could require architectures to dynamically select adaptive forward paths for each pixel rather than for the whole image in [24].

In this paper, we propose the HPS-Net by imitating the hidden markov chain, where an extra mini-branch assists to better select pixel-wise adaptive forward paths in the main-branch for implementing semantic segmentation.

### B. Contextual Information Modeling

In order to capture the contextual information, which implies the relationships between different stuffs, various basic units have been combined in different semantic segmentation structures. Specifically, pooling with different sizes [27] and atrous convolutions with different dilation rates [25], [26], [45] are utilized to incorporate the contextual information of diverse stuffs with different scales. Meanwhile, skip connection structures have been designed to retain features of high resolutions for recovering detailed information [39], [46]. Furthermore, attention mechanisms have been designed in deep structures [47], [48] to dynamically adjust contextual information for every pixel, where spatial attention and channel attention are aggregated. Especially, non-local operations were proposed to obtain dynamic attention maps through pixel-wise similarities [49].

Recently, soft masks have been proposed [29] to adjust the contextual information, which can be seen as the combination of several normalized attention maps to dynamically select adaptive forward paths for every pixel. However, according to the mathematical analyses in terms of parameter optimization we shall provide in this paper, the direct functional relationships between the soft masks and corresponding features in existing pixel-wise dynamic structures [29] would make it difficult to obtain the global optimums.

Consequently, we design an HP-Module to take the advantage of hidden variables derived from the extra mini-branch and adjust the aforementioned deleterious direct relationships, which makes HPS-Net more likely to obtain the optimal pixel-wise path selections.

### C. Benchmark Dataset for Semantic Segmentation

To support developing semantic segmentation algorithms in remote sensing images, benchmark datasets are important to serve as the basis for training and evaluation. Widely used datasets [30], e.g. ISPRS Test Project on Urban Classification and 3D Building Reconstruction, focuses on urban areas in cities such as Vaihingen. Some contests on other topics, such as IEEE GRSS Data Fusion Contest in 2017 and 2018, also provide semantic segmentation datasets [31], [33] covering several cities, while [32] provides data samples from different platforms. Besides, some datasets concentrating on specific categories, such as buildings and roads, are also provided [50]. Further to expand the quantity of data samples, several large scale datasets have been created [34], [35], which cover larger geographical areas to ensure the diversity of data samples. However, existing datasets hardly give considerations to both the quantity of images and sufficiency of land-cover categories. For instance, although GID-5 [34] contains a large quantity of data samples and covers wide ranges of geographical areas, the limited 5 land-cover categories are often insufficient for real-world applications.

In this paper, we expand the GID-5 by refining 5 general categories into 15 elaborated categories and create the GID-15 dataset, which can serve as a challenging benchmark and better meet the demands in real-world applications.
III. Methods

A. Mathematical Formulation

Interpreting the image $I$ of $c \in \mathbb{N}_+$ channels as a mapping function belongs to a special function family $\mathcal{F}$, we have $I : \Omega \to \mathbb{R}^c$ defined on the image grid $\Omega : \{0, 1, \ldots, H - 1\} \times \{0, 1, \ldots, W - 1\}$, which would return values of $c$ channels given a coordinate of position $p$ in $\Omega$. Consequently, the semantic segmentation model can be interpreted as a mapping function that

$$\forall \ I \in \mathcal{F}, \ p \in \Omega, \ f(I, p) = \arg \max_{h \in L} M_1(p).$$

$L = \{l_k\}_{k \in \mathbb{N}_+ \leq N}$ is a label set containing $N$ categories and probability map $M_1$ measures the probability of each category corresponding to each position $p \in \Omega$.

The structures of conventional deep models often combine various elaborate units, such as convolutions and nonlinear activation functions, to capture complex contextual information adaptive to various distributions. Therefore, considering each layer in existing deep structures, such as [39], as a basic function consisting of different units, the whole structures of $D \in \mathbb{N}_+$ layers can be interpreted as a sequence of nested composite functions that

$$\forall \ F_1 = I, \ d \in \mathbb{N}_+ < D, \ F_{d+1} = f_d(\{F_i\}_{i \in \mathbb{N}_+ \leq d}), \ M_1 = F_D, \ f(I, p) = \arg \max_{h \in L} M_1(p).$$

$f_d$ represents the basic function in the $d$-th layer and $F_d$ represents the corresponding feature maps. Considering the skip connection structures, $f_d$ receives feature maps $\{F_i\}_{i \in \mathbb{N}_+ \leq d}$ in all the previous layers (not all necessarily participate in calculations) and returns the feature maps $F_{d+1}$ for $f_{d+1}$ in the next layer. Consequently, the first layer in deep structures receives image $I$ as the input and feature map $F_1$ in the last layer serves as the probability map $M_1$ to measure the pixel-wise probability of each category. It is worth noticing that models derived from NAS can also be interpreted in this form.

In order to better adapt to complicated real-world scenarios, dynamic routing [24] is proposed to select image-dependent forward paths in deep structures. Specifically, we can interpret dynamic routing as

$$\forall \ F_1 = I, \ d \in \mathbb{N}_+ < D, \ F_{d+1} = f_d(\{F_i\}_{i \in \mathbb{N}_+ \leq d}, \alpha^d), \ M_1 = F_D, \ f(I, p) = \arg \max_{h \in L} M_1(p),$$

where $\alpha^d \in \mathbb{R}^s$ represents the activation factor for total $s \in \mathbb{N}_+$ forward paths in the $d$-th layer. In the implementation, dynamic routing utilizes learnable structures to compose the mapping function $G_d : I \to \mathbb{R}^3$ and obtain $\alpha^d$ for selecting forward paths corresponding to up-scale, equivalent-scale and down-scale. The activation factor would be multiplied with corresponding feature maps to select paths adaptive to different scale distributions among the whole input image.

B. Gated Path Selection

Although dynamic routing mentioned in Sec.III-A can adapt to scale distributions during the inference procedure, it rarely considers the diversity over different image areas, which is remarkable in remote sensing images due to the wide range of covered geographical areas. Recently proposed GPSNet [29] extends the idea of dynamic routing, where the utilized soft masks can be seen as the activation factor maps to select adaptive forward paths for each pixel. Specifically, we have

$$\forall \ F_1 = I, d \in \mathbb{N}_+ < D, \ F_{d+1} = f_d(\{F_i \odot W_i^d\}_{i \in \mathbb{N}_+ \leq d}), \ M_1 = F_D, \ f(I, p) = \arg \max_{h \in L} M_1(p).$$

$W_d$ is the soft mask of the same spatial size with feature map $F_i$, where the element in each position serves as a probability for pixel-wise forward path selections. The symbol $\odot$ represents the element-wise product. It is worth noticing that $W_d$ in the traditional layers can be seen as tensors full with 1. In the implementation, GPSNet directly utilizes $F_i$ to generate $W_i^1$, namely $W_i^d = g_i^d(F_i)$ highly dependent with $F_i$ for each $i$ with $g_i^d$ being a differentiable function, which would make it difficult to obtain the global optimums of the problem as we analyze in the following texts.

Noticing that with features in deep structures being considered as variables determined by parameter $\theta_f$ in $\{f_d\}_{d \in \mathbb{N}_+ < D}$, the optimization of $\theta_f$ can be seen as searching the global optimums within a high-dimension manifold $\mathcal{H}_f$ in the full space, whose shape is determined by the structures we design for $\{f_d\}_{d \in \mathbb{N}_+ < D}$. Moreover, optimizing parameters $\theta_f$ in all $\{f_d\}_{d \in \mathbb{N}_+ < D}$ during the training process would guide us to find $F_i^*$ and $W_i^d$ for each $I$ determined by the optimal parameters. Starting from this view, we calculate $\frac{\partial L}{\partial F_i^*}$ and $\frac{\partial L}{\partial W_i^d}$ with $L$ as the loss function such that

$$\frac{\partial L}{\partial F_i^*} = \sum_{d < D} \left(\frac{\partial L}{\partial f_d} \cdot \frac{\partial f_d}{\partial F_i^*} \odot W_i^d\right) \cdot W_i^d,$$

$$\forall d < D, \ \frac{\partial L}{\partial W_i^d} = \frac{\partial L}{\partial f_d} \cdot \frac{\partial f_d}{\partial F_i^*} \odot W_i^d \cdot F_i^*,$$

where $\cdot$ represents the matrix product. Importantly, $\frac{\partial L}{\partial F_i^*}$ and $\frac{\partial L}{\partial W_i^d}$ should be null tensor 0 to retain the probability that
\(F^*_i\) and \(W^*_d\) equal global optimums, which are denoted as \(F^{o}_i\) and \(W^{o}_i\), and \(H_f\) contains the global optimums.

Thus, further with the direct functional relationships \(W^i_d = g^i_d(F_i)\), we require two formulas tenable that

\[
\sum_{d \in D} (\frac{\partial L}{\partial f_d} \cdot \frac{\partial f_d}{\partial F^*_i \circ W^*_i}) \cdot g^i_d(F_i) = \Lambda_i \cdot G_i(\Lambda_{F^*_i}) = 0,
\]

\[
\sum_{d \in D} (\frac{\partial L}{\partial f_d} \cdot \frac{\partial f_d}{\partial F^*_i \circ W^*_i}) \cdot F^*_i = \Lambda_i \cdot \Lambda_{F^*_i} = 0.
\]

(6)

\(\Lambda_i\) and \(\Lambda_{F^*_i}\) are block matrices that list all product factors in the summation on the left side of Eq.(6), while function \(G_i\) is totally determined by \(g^i_d\). Consequently, Eq.(6) illustrates that the column vectors of \(\Lambda_{F^*_i}\) and \(G_i(\Lambda_{F^*_i})\) should be all within the null space of \(\Lambda_i\), which requires numbers of specific points in \(\mathcal{N}_i\) to be mapped within \(\mathcal{N}_i\) through \(G_i\).

In other words, the optimization of \(\theta_f\) can be within a manifold \(\mathcal{H}_{f,g} \subseteq \mathcal{H}_f\), which is narrow down, due to the extra functional relationships \(W^i_d = g^i_d(F_i)\). Then, the extra constraint illustrated above caused by Eq.(6) may not be satisfied by optimal \(\theta_f\), which means the narrow \(\mathcal{H}_{f,g}\) could exclude the global optimums.

Although \(\mathcal{N}_i\) is a linear space, nonlinear units such as softmax or sigmoid are always utilized to fix the ranges of \(G_i(\Lambda_{F^*_i})\). Thus, as illustrated at the top of Fig.3, the designed \(g^i_d\), which determines \(G_i\), is hard to meet the required property (i.e. meeting Eq.(6)) without the prior knowledge of \(\mathcal{N}_i\). Consequently, the narrowed manifold \(\mathcal{H}_{f,g}\) could exclude mentioned global optimums no matter how elaborate the \(g^i_d\) is. The same problem would also exist when \(W^*_d\) has functional relationships with previous layers involved in the calculations of \(F^*_i\), because it also requires \(g^i_d\) to meet the additional property and all restrictions could only be released by optimizing the parameters \(\theta_f\).

C. Hidden Path Selection

In order to release the restriction of \(g^i_d\) caused by direct dependences between \(F_i\) and \(W^i_d\) analyzed in Sec.III-B, we naturally introduce extra variables \(H_d\) highly independent with \(F_i\) to participate in the calculations of \(g^i_d\), namely \(W^i_d = g^i_d(F_i, H_d)\). Serving as the probability mask to indicate the adaptive forward paths for each pixel in \(F_i\), \(W^i_d\) is naturally decided by the properties of different areas in the input image \(I\). Thus, we design the \(H_d\) as the mapping function on \(\mathcal{F}\) defined in Sec.III-A that takes \(I\) as the input. Specifically, we have

\[
W^i_d = g^i_d(F_i, H_d), \quad H_d : \mathcal{F} \rightarrow \mathbb{R}^{s_d},
\]

(7)

where \(s_d\) is a positive integer vector which indicates the shape of \(H_d\). With the help of \(H_d\), the structures of \(g^i_d\) could be of no restrictions to some extent, where \(H_d\) can adjust \(G_i(\Lambda_{F^*_i})\) to approach \(\mathcal{N}_i\) for each \(\Lambda_{F^*_i}\), in \(\mathcal{N}_i\). Equivalently, the shape of manifold \(\mathcal{H}_{f,g}\) would be adjusted by \(H_d\) to approach the global optimums. Consequently, the global optimal with features in deep structures considered as variables are more likely to be searched in the optimization of \(\theta_f\).

Actually, \(H_d\) can be seen as the hidden variables that are unobservable solely given \(F_i\), which inspires us to imitate the hidden markov chain. Specifically, we design a light-weight mini-branch, which generates \(H_d\) to guide forward paths for each pixel of \(F_i\) in the main-branch through Eq.(4) and Eq.(7), namely hidden path selection.

Algorithm 1: Hidden Path Selection Network

\begin{algorithm}
\caption{Hidden Path Selection Network}
\begin{algorithmic}
\STATE Input: image \(I\), \(\mathcal{Z}_d = \{d_k\}_{k \in \mathbb{N}_+}\) for the main-branch, \(\mathcal{H}_d = \{d_m\}_{m \in \mathbb{N}_+}\) for the mini-branch
\STATE Output: Semantic Segmentation Results
\STATE \(1\) \(H_1 = I, F_1 = H_1;\)
\STATE \(2\) for \(d = 1; d < D\) do \(3\) \(H_{d+1} = h_{d}(\{H_i\}_{i \in \mathbb{N}_+ \leq d});\)
\STATE \(4\) for \(d = 1; d < D\) do \(5\) \(d = \min\{d_m \in \mathcal{H}_d, d \leq d_m\};\)
\STATE \(6\) \(d \in \mathcal{Z}_d\) then \(7\) \(F_{d+1} = f_{d}(\{F_i \odot G_i(\Lambda_i)^{-1} \}_{i \in \mathbb{N}_+ \leq d});\)
\STATE \(8\) else \(9\) \(F_{d+1} = f_{d}(\{F_i\}_{i \in \mathbb{N}_+ \leq d});\)
\STATE \(10\) \(M_1 = F_D;\)
\STATE \(11\) final;
\STATE \(12\) return \(\arg \max_{p \in \Omega} M_1(p), \quad \forall \ p \in \Omega;\)
\end{algorithmic}
\end{algorithm}

D. Hidden Path Selection Network

As illustrated in Fig.2, in the proposed HPS-Net, we can implement hidden path selection on existing deep structures (e.g. Resnet-101), which serve as the main branch, based on the analyses in Sec.III-B and Sec.III-C. An extra mini-branch consisting of the same structure as main-branch with all channel numbers reduced to 32 is utilized to generate hidden variables \(H_d\) in Eq.(7). Specifically, we utilize a layer-index set \(\mathcal{Z}_d = \{d_k\}_{k \in \mathbb{N}_+}\) for indicating which layer in the main-branch to be applied hidden path selection. Denoting the basic functions in each layer of mini-branch as \(\{h_d\}_{d \in \mathbb{N}_+ \leq D};\) we also utilize a layer-index set \(\mathcal{H}_d = \{d_m\}_{m \in \mathbb{N}_+}\) to indicate which feature maps in the mini-branch to serve as hidden variables for hidden path selection. Concretely, we modelize the HPS-Net as

\[
\forall \ H_1 = I, \quad \tilde{d} \in \mathbb{N}_+ < D, \quad H_{\tilde{d}+1} = h_{\tilde{d}}(\{H_i\}_{i \in \mathbb{N}_+ \leq \tilde{d}});
\]

\[
\forall \ F_1 = H_1, \quad \tilde{d} \in \mathbb{N}_+ < D, \quad \tilde{d} = \min\{d_m \in \mathcal{H}_d, d \leq d_m\},
\]

\[
F_{\tilde{d}+1} = \begin{cases} f_{\tilde{d}}(\{F_i \odot G_i(\Lambda_i)^{-1}\}_{i \in \mathbb{N}_+ \leq \tilde{d}}), & \text{if } \tilde{d} \in \mathcal{Z}_d, \\ f_{\tilde{d}}(\{F_i\}_{i \in \mathbb{N}_+ \leq \tilde{d}}), & \text{otherwise}, \end{cases}
\]

\[
M_1 = F_D, \quad f(I, p) = \arg \max_{\tilde{d} \in \mathcal{B}} M_1(p),
\]

(8)

where the output \(F_D\) of main-branch finally implements semantic segmentation. The detail process can be found in Algorithm 1.

In the implementation, as illustrated in Fig.4, we propose Hidden Path Module (HP-Module) to generate \(W^i_d = g^i_d(F_i, H_i)\), which consists of feature concatenations, convolutions and normalization functions. We have

\[
W^i_d = \lceil p_{\text{norm}}(f_{\text{cat}}(f_{\text{conv}}(F_i, H_i))) \rceil_i.
\]

(9)
where \( f_{\text{conv}_2} \) and \( f_{\text{conv}_32} \) are convolutions with 2 and 32 output channels respectively, where kernel sizes are all set as \( 3 \times 3 \). \( f_{\text{out}} \) is the concatenation operation. \( f_{\text{norm}} \) is the normalization function consisting of softmax and clip function, which normalizes the values within \([\alpha, \beta]\). Importantly, \([\cdot, \cdot]_i\) indicates the slice alongside the channel dimension according to index \( i \), which is corresponding to the \( i \)-th alternative path.

Recalling analyses on the global optimums in Sec.III-B, we further calculate the gradient flows of \( F_i \) with \( W_i^d = g_i(\hat{F}_i, \hat{H}_i) \) that

\[
\frac{\partial L}{\partial F_i} = \sum_{d \in \mathbb{D}} \frac{\partial L}{\partial f_d} \cdot \frac{\partial f_d}{\partial F_i} \cdot (W_i^d + F_i \cdot \frac{\partial W_i^d}{\partial F_i}). \tag{10}
\]

It is worth noticing that Eq.(5) and Eq.(10) are not contradicted. Eq.(5) calculates the partial derivative of \( F_i \) with features in deep structures (e.g. \( W_i^d \) and \( F_i \)) considered as variables, which is utilized to analyze the global optimums in the actual full space and not involved in the calculations among parameter optimization. Eq.(10) calculates the gradient flows in our proposed HPS-Net to optimize the parameter \( \theta_f \), which can be seen as searching parameters within a manifold \( \mathcal{H}_{fh} \) adjusted by \( H_d \) based on \( \mathcal{H}_{fg} \). Moreover, the efficiency of training process would be increased with \( W_i^d \) being contrary sign to \( F_i \cdot \frac{\partial W_i^d}{\partial F_i} \), which is hard to avoid with constantly changing parameters \( \theta_f \) during the training process. Consequently, we cut-off the gradient flows from soft masks \( W_i^d \) to feature maps \( F_i \) in the main-branch by multiplying null tensor \( 0 \) with the \( F_i \cdot \frac{\partial W_i^d}{\partial F_i} \) on the right side of Eq.(10). As we shall see, by controlling gradient flows, the efficiency of training process would be increased.

E. Model Analyses

To hark back to the manifold for parameter optimization, as illustrated at the bottom of Fig.3, the shape of \( \mathcal{H}_{fh} \) could be adjusted by \( H_d \) based on \( \mathcal{H}_{fg} \), where we realize that the high independence between \( F_i \) and \( H_d \) makes it possible to let \( H_d \) value freely and adjust \( W_i^d \) (i.e. adjust \( G_i(A_{\hat{F}_i}) \)) to meet the demanded property. Consequently, the larger range domain of \( H_d \) in the mini-branch would make \( \mathcal{H}_{fh} \) adjusted more freely with the designed formulation of \( g_i^d \), where \( \mathcal{H}_{fh} \) is more likely to contain the global optimums with all feature maps considered as variables.

Moreover, the floating range of values in \( W_i^d \) is defined within \([\alpha, \beta]\), which is ensured by applying \( f_{\text{norm}} \). Recalling that we denote the global optimums \( W_i^d \) as \( W_i^d^* \), we focus on the Taylor expansion with respect to variable \( W_i^d \) that

\[
L = L(W_i^d^*) + (W_i^d - W_i^d^*)^T \cdot \frac{\partial L}{\partial W_i^d} \\
+ \frac{1}{2} (W_i^d - W_i^d^*)^T \cdot \frac{\partial^2 L}{\partial (W_i^d)^2} \cdot (W_i^d - W_i^d^*) + o(||W_i^d - W_i^d^*||), \tag{11}
\]

where \( o(\cdot) \) represents the higher order infinitesimal. Since we fix the range of \( W_i^d \) within \([\alpha, \beta]\), the expectation values of \( |W_i^d - W_i^d^*| \) is \((\beta - \alpha)/2\), which is 0.25 in the implementation. Noticing that \( \frac{\partial^2 L}{\partial (W_i^d)^2} \) is a null matrix, the second order term in the Taylor expansion would dominate the calculations. Then, with \( \frac{\partial^2 L}{\partial (W_i^d)^2} \) being a positive definite matrix, the loss function \( L \) approximates a convex function to some extent with respect to \( W_i^d \). The convexity would make it easier to search the global optimums with freely valuing \( W_i^d \) by adjusting \( H_d \) if \( \mathcal{H}_{fh} \) contains the global optimums.

Consequently, HPS-Net could provide more probability to obtain promising results due to the better manifold for parameter optimization, which can make HPS-Net obtain better pixel-wise path selections and better depict pixel-wise land-cover distributions during the inference procedure. However, compared with parameter optimization of conventional deep structures without path selections, Eq.(10) demonstrates that gradient flows of HPS-Net are considerable small due to the extra multiplication of \( W_i^d \), which might make hidden path selection more effective under small training costs compared with larger training costs.

F. Loss Function

We utilize fully supervised learning to optimize parameters \( \theta_f \) in the proposed HPS-Net. Denoting \( G_i \) as the pixel-wise ground truth of \( I_i \), which indicates the semantic category for each pixel, we have

\[
L = \text{Cross-Entropy}(M_i, G_i), \tag{12}
\]

where the cross entropy function takes the model output \( M_i \) in Eq.(8) and ground truth \( G_i \) as the input. Stochastic gradient descent (SGD) is then used in the optimization process of parameter \( \theta_f \), through which the mini-branch and main-branch can be trained simultaneously.

IV. THE GID-15 DATASET

As the basis for training and evaluation, a well-annotated benchmark dataset is important to develop remote sensing semantic segmentation algorithms. Although some widely used datasets with several land-cover categories [30], [31], [33], [50] are provided, remote sensing semantic segmentation datasets are hard to give considerations to both the quantities of data samples and sufficiencies of land-cover categories. As a consequence, the increasing demands in real-world applications, such as identifying industrial land and urban residential simultaneously, are difficult to satisfy. For instance, containing abundant data samples that cover dozens of cities in China with 150 images of 6800 × 7200 pixels, GID-5 [34] ensures
Fig. 5. Data samples taken from our proposed GID-15 dataset. Different colors indicate different land-cover categories. The obscure land covers are blackened in the annotation masks. Several intact data samples are listed at the top and some cropped data patches are listed on the bottom, where we can see the annotation meticulousness with wide geographical areas covered.

the data diversity for deep model training, while the only 5 involved land-cover categories (i.e. built-up, farmland, forest, meadow and water) are not sufficient for real applications. In order to propose a better benchmark meeting practical demands, we expand the fine land-cover classification set in [3] by subdividing 5 land-cover categories in the GID-5 to compose a new GID-15.

Specifically, to further distinguish sub-categories belonging to the same general land-cover categories, we refer to the Chinese Land Use Classification Criteria (GB/T21010-2017) and finally determine 15 concerned categories through a hierarchical category relationships. Concretely, for researching the detailed urban area distributions, we subdivide the built-up into industrial land, urban residential, rural residential and traffic land. Then, the forest is subdivided into garden land, arbor forest and shrub land for the researches on vegetation covers, while the meadow is subdivided into natural meadow and artificial meadow. Last but not the least, the farmland is subdivided into paddy field, irrigated land and dry cropland for researching agricultural land distributions, while the water is subdivided into river, lake and pond for the water resources researching.

In practice, an expert group checks the original images and annotated category maps in GID-5 simultaneously, through which we recheck the previous annotations and further separate the annotated land-cover regions in GID-5 according to the hierarchical category relationships we build. Similar with GID-5, all the pixels related to obscure land-cover distributions, which are difficult to depict with pixel-wise land-cover categories explicitly, are ignored in GID-15. Further due to the wide range areas covered in GID-5, GID-15 can serve as a challenging benchmark containing large quantities of data samples (i.e. 150 images of 6800 × 7200 pixels with pixel-wise annotations) and abundant land-cover categories (i.e. 15 land-cover categories better meeting practical demands) simultaneously. Several samples of GID-15 are displayed in Fig.5, where we can see the annotation meticulousness. The GID-15 dataset would soon be available at https://captain-whu.github.io/HPS-Net/.

V. EXPERIMENTS AND ANALYSIS

In this section, we evaluate the proposed HPS-Net on both GID-5 and GID-15, where we show the effectiveness of the proposed modules and verify the mathematical analyses in Sec.III. We first illustrate the evaluation metric in Sec.V-A, while we further clarify the detail experiment settings in Sec.V-B. Then, in Sec.V-C and Sec.V-D, we discuss the application of the proposed structures on different classical deep
structures [39] and the state-of-the-art pixel-wise dynamic algorithm [29], where we explore the merits of proposed hidden path selection. Further in Sec.V-E, we present detail verifications on the mathematical analyses in Sec.III and study the effects of each term in the proposed structures through ablation studies and feature visualizations.

A. Evaluation Metrics

To reduce the influence of label imbalance, we follow the principle in [29] that utilize mean Intersection over Union (mIoU). Denoting the confusion matrix calculated based on $M_I$ and $G_I$ as $Q = \{q_{ij}\}$, we have

$$mIoU = \frac{1}{N} \sum_{i=1}^{N} q_{ii} / \left[ \sum_{j=1}^{N} (q_{ij} + q_{ji}) - q_{ii} \right],$$  \hspace{1cm} (13)$$

where $N$ is the total number of land-cover categories. With each term on the right side of Eq.(13) measuring the identification of one certain category, the average form ensures the equivalence of each category.

B. Experiment Settings

The deep structures and semantic segmentation algorithms involved in our experiments are as follows:

- Resnet-50 [39]: a deep residual structure serving as the main branch in HPS-Net, where 16 skip connections are considered to be applied hidden path selection.
- Resnet-101 [39]: a deeper version of residual structure compared with Resnet-50, where 33 skip connections are considered to be applied hidden path selection.
- Pspnet [27]: a semantic segmentation algorithm that utilizes multi-scale structures based on poolings of different size.
- DeepLabv3* [26]: a semantic segmentation algorithm taking residual structures as the backbone, which integrates the encoder-decoder structures and multi-scale structures based on dilation convolutions.
- GPSNet [29]: a semantic segmentation algorithm that dynamically selects adaptive forward paths for each pixel.

In this paper, we choose residual structures as the main branch in HPS-Net. All the experiments are implemented on a Pascal V100 with 16G memory based on Pytorch. All the model training and evaluation are under the same conditions without pre-trained parameters and post-processings. In the training process, SGD is utilized with batch size as 10 and random flip is utilized as the data augmentation. The initial learning rate is set as 0.007 and poly policy is employed with the power of 0.9. We set the momentum as 0.9 and weight decay as 0.0001. In the testing process, we take raw outputs of each model as the results for the evaluation.

As for the hyper-parameters in HPS-Net, we set $\mathcal{L}_d = \{d_k\}_{k \in \mathbb{N}_+}$ as the index of layers containing skip connections, which indicates that total 16 layers in Resnet-50 and 33 layers in Resnet-101 are applied hidden path selection in the main-branch. Then, with residual structures divided into several stages according to the pixel resolution, we set

$$\mathcal{H}_d = \{d_m\}_{m \in \mathbb{N}_+}$$

as the index of last layer in each stage, which indicates that total 4 layers serve as the hidden variables derived from the mini-branch. Finally, we generally set $(\alpha, \beta)$ as $(0.75, 1.25)$ in the HP-Module, where the softmax function is multiplied with 2 to avoid the gradient vanishing. Especially, in the first layer of each stage in the main-branch, $(\alpha, \beta)$ are expanded to $(0.5, 1.5)$ due to the additional alternative paths from previous stages.

For all the experiments in both GID-5 and GID-15, we crop the images and corresponding annotation masks into patches of $512 \times 512$ pixels simultaneously, which are split into two subsets, i.e. totally 25200 data patches for training and 6300 data patches for testing.

C. Discussions on Hidden Path Selection

For simply discussing the application of hidden path selection on existing deep structures, we first run a toy experiment, where 10% data samples are randomly selected from GID-15 for training and evaluation. Specifically, we apply the hidden path selection on Resnet-101 in DeepLabv3* to compose HPS-Net, which is utilized to compare with the original residual structure. The experimental results with different epochs, i.e. from 10 epochs to 100 epochs at 10 intervals, are shown in Fig.6, where we can see the fold lines marking evaluated mIoU corresponding to each model with the same Resnet-101 backbone.

Specifically, as for the fold line corresponding to Deeplabv3*, we can see that the model encounters a bottleneck around 32.0 in mIoU. However, with only 10% data samples selected, the model training is far from saturation. Thus, this bottleneck indicates that the parameter optimization in DeepLabv3* on the high dimension manifold hovers at the same level, which implies the defects of aforementioned manifold. Then, for the fold line corresponding to HPS-Net, the model performance is constantly improved with increasing training epochs, which implies the better shape of manifold for parameter optimization in HPS-Net to make global optimums
Fig. 7. Examples of comparing our HPS-Net with the state-of-the-art pixel-wise dynamic method GPSNet on GID-15 with Resnet-101 as the backbone. Our proposed HPS-Net can better depict pixel-wise land-cover distributions. The blackened pixels are the ignored regions in the annotation masks, which are also ignored in the semantic segmentation results.

TABLE I

COMPARISONS WITH STATE-OF-THE-ART METHODS TRAINED ON GID-15. WE DENOTE APPLYING THE HIDDEN PATH SELECTION ON Deeplabv3* AS HPS-Net-101 and HPS-Net-50, WHILE APPLYING HIDDEN PATH SELECTION ON GPSNet WITH DIFFERENT BACKBONES ARE DENOTED AS HPS-Net-G101 and HPS-Net-G50. NUMBER 50 AND NUMBER 101 FOLLOWING THE MODEL NAME INDICATE THE RESNET-50 AND RESNET-101 BACKBONES. LAND-COVER CATEGORIES LISTED IN THE FIRST ROW ARE ABBREVIATIONS FOLLOWING THE ORDER IN FIG. 5.

| Methods                  | P. field | Irr. land | Dry crop. | Garden | Arb. forest | Shr. land | Nat. mead. | Art. mead. | Ind. land | Urb. resid. | Rur. resid. | Traff. land | River | Lake | Pond | mIoU |
|--------------------------|----------|-----------|-----------|---------|-------------|-----------|------------|------------|-----------|-------------|-------------|------------|--------|------|------|------|
| PSPNet [27]              | 54.4     | 74.3      | 40.1      | 21.4    | 85.6        | 11.9      | 68.4       | 30.4       | 62.9      | 73.4        | 61.4        | 54.9      | 57.5  | 68.0 | 28.7 | 52.9 |
| Deeplabv3*-50 [26]       | 56.6     | 70.2      | 41.1      | 21.2    | 89.2        | 15.6      | 64.5       | 37.9       | 64.1      | 73.9        | 62.5        | 59.3      | 60.9  | 73.5 | 23.6 | 54.3 |
| Deeplabv3*-101 [26]      | 55.7     | 74.2      | 44.0      | 23.5    | 89.3        | 12.2      | 64.0       | 38.9       | 64.4      | 74.4        | 63.7        | 57.6      | 55.2  | 70.7 | 27.9 | 54.1 |
| GPSNet-50 [29]           | 54.6     | 72.3      | 41.1      | 23.1    | 90.1        | 15.6      | 66.5       | 38.8       | 63.6      | 74.1        | 63.9        | 57.1      | 55.6  | 71.9 | 26.7 | 54.3 |
| GPSNet-101 [29]          | 53.4     | 73.4      | 43.8      | 16.8    | 88.8        | 16.5      | 66.2       | 42.0       | 63.2      | 73.4        | 61.2        | 56.7      | 52.9  | 71.5 | 25.6 | 53.7 |
| HPS-Net-50 (Ours)        | 54.1     | 73.4      | 41.9      | 24.3    | 90.1        | 13.2      | 66.7       | 40.7       | 65.2      | 75.0        | 65.2        | 60.8      | 57.1  | 71.4 | 26.4 | 55.0 |
| HPS-Net-101 (Ours)       | 55.5     | 74.7      | 47.6      | 20.8    | 88.7        | 12.7      | 63.9       | 41.3       | 64.7      | 74.6        | 63.1        | 59.5      | 58.9  | 71.5 | 25.8 | 54.9 |
| HPS-Net-G50 (Ours)       | 56.1     | 74.0      | 47.0      | 20.4    | 91.0        | 17.3      | 68.8       | 44.0       | 63.9      | 74.7        | 65.2        | 59.7      | 52.1  | 71.5 | 26.5 | 55.5 |
| HPS-Net-G101 (Ours)      | 53.2     | 75.6      | 47.3      | 22.1    | 91.1        | 17.0      | 69.5       | 45.2       | 63.3      | 74.0        | 64.4        | 59.0      | 55.6  | 72.4 | 23.4 | 55.5 |

more accessible as demonstrated by the mathematical analyses in Sec.III-E. As for the comparison of two fold lines, HPS-Net outperforms Deeplabv3* by average 3.2 in mIoU, which indicates the improvement attributed to applying hidden path selection on the basic residual structure.

D. Comparison with the State-of-the-art

To compare with state-of-the-art algorithms, we list the IoU for each land-cover category and mIoU for overall evaluation in Tab.I, where all models are trained on GID-15 with 20 epochs from scratch. We apply the hidden path selection on Deeplabv3* and GPSNet to compose the HPS-Net. We denote applying the hidden path selection on Deeplabv3* as HPS-Net-50 and HPS-Net-101, while applying hidden path selection on GPSNet with different backbones are denoted as HPS-Net-G50 and HPS-Net-G101. Number 50 and number 101 following the model name indicate the Resnet-50 and Resnet-101 backbones. As we can see, HPS-Net tends to achieve better results on the majority of land-cover categories, while the categories related to water can not be depicted that well. Furthermore we can conclude from Tab.I that HPS-Net-G50 and HPS-Net-G101 achieve the best results, i.e. 55.5 in mIoU.
specifically, HPS-Net-50 outperforms Deeplabv3*-50 by 0.7 in mIoU and HPS-Net-101 raises 0.8 improvement over Deeplabv3*-101, while HPS-Net-G50 outperforms GPSNet-50 by 1.2 in mIoU and HPS-Net-G101 raises 1.8 improvement over GPSNet-101. The model comparisons listed above illustrate that the applied hidden path selection can stably improve the performance of existing deep structures. Besides, we can check that GPSNet-50 and GPSNet-101 achieve comparative results with Deeplabv3*-50 and Deeplabv3*-101, which indicates that the proposed hidden path selection can build better pixel-wise dynamic structures than the gate path selection proposed in the state-of-the-art algorithm GPSNet. However, together with the application of hidden path selection, gate path selection is able to raise more stable improvements according to the comparison between HPS-Net-50 and HPS-Net-G50 or HPS-Net-101 and HPS-Net-G101. We display the semantic segmentation results in Fig. 7, where we can also see the superiorities of the proposed HPS-Net.

For the experiments on GID-5, we recover the 5 land-cover categories in GID-5 from the data samples involved in experiments mentioned above according to the hierarchical category relationships in Sec.IV, based on which all models are also trained for 20 epochs from scratch. The experimental results are summarized in Tab.II, where we can also see that HPS-Net obtains better results. However, it is worth noticing that, the gaps between HPS-Net and other algorithms are smaller than that in the experiment results on GID-15, which demonstrates that GID-15 can better evaluate different semantic segmentation algorithms. We also illustrate the semantic segmentation results in Fig. 8 to show the merits of applying hidden path selection on the basic residual structures.

E. Ablation Study and Feature Visualization

To study the effects of each term in the proposed structures, we design several ablation studies in this section. Firstly, we average the mask $W^d_i$ in HPS-Net and make all pixels sharing the same path selection, denoted as HPS-Net-50-ps and HPS-Net-101-ps. Subsequently, we can explore the merits of pixel-wise dynamic structures compared with image-wise dynamic structures. Then, we fix the hidden variables in HPS-Net, denoted as HPS-Net-50-fh and HPS-Net-101-fh, to study the influence of direct relationships between $W^d_i$ and $F_i$ analyzed in Sec.III-B. Besides, we further keep intact gradient flows in
Fig. 9. Visualization of the masks corresponding to path selection in HPS-Net-G101 and GPSNet-101. We visualize the path selection for the first layer of 1-st stage to 3-rd stage in HPS-Net-G101, while the first mask of 1-st branch to 3-rd branch in GPSNet are listed too. The brighter grayscale values in these masks indicate larger factors corresponding to the first alternative path, where the proposed HPS-Net can better depict pixel-wise land-cover distributions through diverse forward path selections for different pixels compared with GPSNet.

### TABLE III

| Methods          | P. field | Irr. land | Dry. crop | Garden | Arb. forest | Shr. land | Nat. mead. | Arb. mead | Ind. land | Urb. resid. | Rur. resid. | Traff. land | River | Lake | Pond | mIoU |
|------------------|----------|-----------|-----------|--------|-------------|-----------|------------|-----------|-----------|-------------|------------|-------------|-------|------|------|------|
| HPS-Net-50-ps    | 52.5     | 72.9      | 42.0      | 24.4   | 15.3        | 64.1      | 40.7       | 64.8      | 74.9      | 64.5        | 60.2       | 59.5        | 69.6  | 24.3 | 54.5 |
| HPS-Net-101-ps   | 54.4     | 72.8      | 46.3      | 19.4   | 99.9        | 61.5      | 33.6       | 59.4      | 73.6      | 61.3        | 55.4       | 50.7        | 68.2  | 29.2 | 52.7 |
| HPS-Net-50-fh    | 52.0     | 71.8      | 37.9      | 22.8   | 12.8        | 63.1      | 39.2       | 63.9      | 74.3      | 63.8        | 59.9       | 53.8        | 65.5  | 22.3 | 52.6 |
| HPS-Net-101-fh   | 57.1     | 72.4      | 40.6      | 20.6   | 12.5        | 62.9      | 34.7       | 62.8      | 73.6      | 63.5        | 57.6       | 48.9        | 64.6  | 22.7 | 52.1 |
| HPS-Net-50-ig    | 55.9     | 72.2      | 38.0      | 20.6   | 13.0        | 63.0      | 38.9       | 63.8      | 74.0      | 63.6        | 58.4       | 57.4        | 70.3  | 25.2 | 53.6 |
| HPS-Net-101-ig   | 55.9     | 72.2      | 38.0      | 20.6   | 13.0        | 63.0      | 38.9       | 63.8      | 74.0      | 63.6        | 58.4       | 57.4        | 70.3  | 25.2 | 53.1 |
| HPS-Net-50       | 54.1     | 73.4      | 41.9      | 24.3   | 13.2        | 66.7      | 40.7       | 65.2      | 75.0      | 65.2        | 60.8       | 57.1        | 71.4  | 26.4 | 55.0 |
| HPS-Net-101      | 55.2     | 74.7      | 47.6      | 20.8   | 88.7        | 63.9      | 41.3       | 64.7      | 74.6      | 63.1        | 59.5       | 58.9        | 71.5  | 25.8 | 54.9 |

As illustrated in Tab.III, HPS-Net-50-ps and HPS-Net-101-ps are inferior to HPS-Net-50 and HPS-Net-101 by 0.5 and 2.2 in mIoU respectively, which indicates the superiority of pixel-wise dynamic structures. Then, HPS-Net-50 and HPS-Net-101 outperform HPS-Net-50-fh and HPS-Net-101-fh by 2.4 and 2.8 in mIoU respectively, which verifies the deleteriousness of direct relationships between $W_d$ and $F_i$ analyzed in Sec.III-B. Finally, HPS-Net-50-ig and HPS-Net-101-ig achieve lower results by 1.4 and 1.8 in mIoU than HPS-Net-50 and HPS-Net-101 respectively, which verifies the inefficiency of model training with intact gradient flows analyzed in Sec.III-D.

For illustrating the depiction of pixel-wise land-cover distributions by HPS-Net during the inference procedure, we visualize the selected paths of several data samples in Fig.9, where the proposed HPS-Net can better depict pixel-wise land-cover distributions through dynamic pixel-wise forward paths compared with GPSNet. Specifically, as illustrated in the first row of Fig.9, HPS-Net emphasizes paths with more complex computations, i.e. the first alternative path, among rural residential at top of the image, which alleviates the mis-prediction compared with GPSNet. Moreover, as shown in the second row, the combination of path selections from 1-st stage to 3-rd stage ensures different forward paths for irrigated land at left-top and dry cropland at right-bottom, which makes efforts to distinguish these two land-cover categories in the same image and also alleviate the mis-prediction. As we can see, the proposed HPS-Net can better depict pixel-wise land-cover distributions during the inference procedure.

### TABLE IV

| Backbones | Deeplabv3* | HPS-Net-w/o-h | HPS-Net |
|-----------|------------|---------------|---------|
| Resnet-50 | 69.2       | 74.9          | 75.2    |
| Resnet-101| 88.6       | 94.6          | 95.0    |

the HP-Module, denoted as HPS-Net-50-ig and HPS-Net-101-ig, to study the gradient analyses in Sec.III-D.

As illustrated in Tab.III, HPS-Net-50-ps and HPS-Net-101-ps are inferior to HPS-Net-50 and HPS-Net-101 by 0.5 and 2.2 in mIoU respectively, which indicates the superiority of pixel-wise dynamic structures. Then, HPS-Net-50 and HPS-Net-101 outperform HPS-Net-50-fh and HPS-Net-101-fh by 2.4 and 2.8 in mIoU respectively, which verifies the deleteriousness of direct relationships between $W_d$ and $F_i$ analyzed in Sec.III-B. Finally, HPS-Net-50-ig and HPS-Net-101-ig achieve lower results by 1.4 and 1.8 in mIoU than HPS-Net-50 and HPS-Net-101 respectively, which verifies the inefficiency of model training with intact gradient flows analyzed in Sec.III-D.

For illustrating the depiction of pixel-wise land-cover distributions by HPS-Net during the inference procedure, we visualize the selected paths of several data samples in Fig.9, where the proposed HPS-Net can better depict pixel-wise land-cover distributions through dynamic pixel-wise forward paths compared with GPSNet. Specifically, as illustrated in
In this paper, we propose a hidden path selection network for semantic segmentation in remote sensing images to select adaptive forward paths for every pixel guided by the mathematical analyses in terms of the parameter optimization. The inherent problem about inaccessible global optimums is tackled with the help of hidden variables derived from an extra mini-branch in HPS-Net. For the better training and evaluation, we expand the land-cover classification dataset, i.e. GID-5, into 15 land-cover categories and propose the new GID-15 dataset, which provides more abundant land-cover categories and better satisfies the real-world scenarios. The experimental results on both GID-5 and GID-15 show that the proposed module can stably improve the performances of existing deep structures.

VI. CONCLUSION

Finally, to explore the efficiency of proposed HPS-Net, we list the computation costs of each model in Tab.IV. As we can see, pixel-wise dynamic structures cause 5.7G extra FLOPs and hidden variables result in 0.3G extra FLOPs with ResNet-50 as the backbone, while the corresponding extra FLOPs are 6.0G and 0.4G respectively with ResNet-101 as the backbone. Compared with the FLOPs of Deeplabv3*, HPS-Net results in less than 10% extra computation costs, which indicates that HPS-Net gives the consideration to both effectiveness and efficiency.

REFERENCES

[1] C. Zhang, S. Isabel, P. Xin, H. Li, G. Andy, H. Jonathon, and P. M. Atkinson, “An object-based convolutional neural network (ocnn) for urban land use classification,” Remote Sens. Environ., vol. 216, pp. 57–70, 2018.

[2] C. Zhang, I. Sargent, P. Xin, H. Li, and P. M. Atkinson, “Joint deep learning for land cover and land use classification,” Remote Sens. Environ., vol. 221, pp. 173–187, 2019.

[3] X. Y. Tong, G. S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, “Land-cover classification with high-resolution remote sensing images using transferable deep models,” Remote Sens. Environ., vol. 237, p. 111322, 2020.

[4] F. Ybue, A. Gs, L. A. Yu, B. Pm, L. C. Gang, and D. Yz, “Comprehensively analyzing optical and polarimetric sar features for land-use/land-cover classification and urban vegetation extraction in highly-dense urban area,” Int. J. Appl. Earth Obs., vol. 103.

[5] J. F. Fekel, A. Cottam, N. Gorelick, and A. S. Belward, “High-resolution mapping of global surface water and its long-term changes,” Nature.

[6] K. Pollock, “Policy: Urban physics,” Nature, vol. 531, no. 7594, p. S64, 2016.

[7] Arneth and Almut, “Climate science: Uncertain future for vegetation cover,” Nature, vol. 524, no. 7563, pp. 44–5, 2015.

[8] G. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, “AID: A benchmark data set for performance evaluation of aerial scene classification,” IEEE Trans. Geosci. Remote Sens., vol. 55, no. 7, pp. 3965–3981, 2017.

[9] Y. Long, G. Xia, S. Li, W. Yang, M. Y. Yang, X. X. Zhu, L. Zhang, and D. Li, “On creating benchmark dataset for aerial image interpretation: Reviews, guidelines, and million-aid,” IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 14, pp. 4205–4230, 2021.

[10] A. Ghosh, B. N. Subudhi, and L. Bruzzone, “Integration of gibbs markov random field and hopfield-type neural networks for unsupervised change detection in remotely sensed multi-temporal images,” IEEE Trans. Image Process., vol. 22, no. 8, pp. 3087–3096, 2013.

[11] G. S. Xia, X. Bai, J. Ding, Z. Zhu, S. J. Belongie, J. Luo, M. Dutc, M. Pelillo, and L. Zhang, “DOTA: A large-scale dataset for object detection in aerial images,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 3974–3983.

[12] J. Ding, N. Xue, G. Xia, X. Bai, W. Yang, M. Y. Yang, S. J. Belongie, J. Luo, M. Dutc, M. Pelillo, and L. Zhang, “Object detection in aerial images: A large-scale benchmark and challenges,” IEEE Trans. Pattern Anal. Mach. Intell.

[13] B. Du, R. Lu, C. Wu, and L. Zhang, “Unsupervised deep slow feature analysis for change detection in multi-temporal remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 57, no. 12, pp. 9876–9992, 2019.

[14] C. Wu, B. Du, and L. Zhang, “Slow feature analysis for change detection in multispectral imagery,” IEEE Trans. Geosci. Remote Sens., vol. 52, no. 5, pp. 2858–2874, 2014.

[15] C. Zhang, P. A. Harrison, X. Pan, H. Li, and P. M. Atkinson, “Scale sequence joint deep learning (ss-jdl) for land use and land cover classification,” Remote Sens. Environ., vol. 237, pp. 1–16, 2020.

[16] A. Vsm, A. Alk, A. Bkg, B. Hlfds, and C. Caa, “Exploring multiscale object-based convolutional neural network (multi-ocnn) for remote sensing image classification at high spatial resolution - scenedirect,” ISPRS J. Photogrammetry Remote Sens., vol. 168, pp. 56–73, 2020.

[17] E. A. Alshari and B. W. Gawali, “Development of classification system for lule using remote sensing and gis,” Glob. Transitions Proc., 2021.

[18] J. Mingers, “An empirical comparison of selection measures for decision-tree induction,” Machine Learning, 1989.

[19] D. V. Lindley, “Regression lines and the linear functional relationship,” J. Roy. Statist. Soc. Suppl., vol. 9, no. 2, 1947.

[20] M. Xu, P. Watanachaturaporn, P. K. Varshney, and M. K. Arora, “Decision tree regression for soft classification of remote sensing data,” Remote Sens. Environ., 2005.

[21] E. Lopez, G. Bocco, M. Mendoza, and E. Duhan, “Predicting land-cover and land-use change in the urban fringe: A case in morelia city, mexico,” Landsca. Urban Plann., vol. 55, no. 4, pp. 271–285, 2001.

[22] A. Vali, S. Comai, and M. Matteucci, “Deep learning for land use and land cover classification based on hyperspectral and multispectral earth observation data: A review,” Remote Sens., vol. 12, no. 15, p. 2495, 2020.

[23] K. Yang, Z. Liu, Q. Lu, and G. Xia, “Multi-scale weighted branch network for remote sensing image classification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, 2019, pp. 1–10.

[24] Y. Li, L. Song, Y. Chen, Z. Li, X. Zhang, X. Wang, and J. Sun, “Learning dynamic routing for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 8550–8559.

[25] L. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” CoRR, vol. abs/1706.05587, 2017.

[26] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 833–851.

[27] H. Zhao, I. Shi, X. Qi, X. Wang, and I. Jia, “Pyramid scene parsing network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 6230–6239.

[28] C. Liu, L. Chen, F. Schroff, H. Adam, W. Hua, A. L. Yuille, and L. Fei-Fei, “Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation,” CoRR, vol.abs/1901.02985, 2019.

[29] Q. Geng, H. Zhang, X. Qi, G. Huang, R. Yang, and Z. Zhou, “Gated path selection network for semantic segmentation,” IEEE Trans. Image Process., vol. 30, pp. 2436–2449, 2021.

[30] F. Rottensteiner, G. Sohn, J. Jung, M. Gerke, and U. Breitkopf, “The isprs benchmark on urban object classification and 3d building reconstruction,” ISPRS Ann. of Photogramm. Remote Sens. Spatial Inf. Sci., vol. 12.

[31] Denis, Tuia, Gabriele, Moser, Bertrand, Le, Saux, Benjamin, Bechtel, and Linda, “2017 ieee gis data fusion contest: Open data for global multimodal land use classification [technical committees],” IEEE Geosci. Remote Sens. Mag., 2017.

[32] M. Zhang, X. Hu, L. Zhao, Y. Lv, M. Luo, and S. Pang, “Learning dual multi-scale manifold ranking for semantic segmentation of high-resolution images,” Remote Sens., vol. 9, no. 5, p. 500, 2017.

[33] B. L. Saux, N. Yokoya, R. Hansch, and S. Prasad, “2018 ieee gis data fusion contest: Multimodal land use classification [technical committees],” IEEE Geosci. Remote Sens. Mag., vol. 6, no. 1, pp. 52–54, 2018.

[34] X. Tong, Q. Lu, G. Xia, and L. Zhang, “Large-scale land cover classification in gaofen-2 satellite imagery,” in Proc IEEE Int. Geosci. Remote Sens. Symp., 2018, pp. 3599–3602.

[35] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, “Deepglobe 2018: A challenge to parse the earth through satellite images,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, June 2018.
[36] Ke, Zou, Changqing, Shuhui, Liang, Yun, Zhang, Jian, Gong, and Minglun, “Multi-modal feature fusion for geographic image annotation,” Pattern Recognit., vol. 73, pp. 1–14, 2018.

[37] Tuia, Devis, Pacifici, Fabio, Kanevski, Mikhail, Emery, and J. William, “Classification of very high spatial resolution imagery using mathematical morphology and support vector machines.” IEEE Trans. Geosci. Remote Sens., 2009.

[38] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. Int. Conf. Learn. Representations, 2015.

[39] S. Xie, R. B. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 5987–5995.

[40] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions.” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., June 2015.

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in NeurIPS, 2012, pp. 1097–1105.

[42] E. Shelhamer, J. Long, and T. Darrell, “Fully convolutional networks for semantic segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 4, pp. 640–651, 2017.

[43] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 12, pp. 2481–2495, 2017.

[44] H. Noh, S. Hong, and B. Han, “Learning deconvolution network for semantic segmentation,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 1520–1528.

[45] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” CoRR, vol. abs/1511.07122, 2015.

[46] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in MICCAI, 2015, pp. 234–241.

[47] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, “Asymmetric non-local neural networks for semantic segmentation,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 593–602.

[48] H. Li, P. Xiong, J. An, and L. Wang, “Pyramid attention network for semantic segmentation,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 593–602.

[49] X. Wang, R. B. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 7794–7803.

[50] P. Kaiser, J. D. Wegner, A. Lucchi, M. Jaggi, T. Hofmann, and K. Schindler, “Learning aerial image segmentation from online maps,” IEEE Trans. Geosci. Remote Sens., vol. 55, no. 11, pp. 6054–6068, 2017.