

TriTransNet: RGB-D Salient Object Detection with a Triplet Transformer Embedding Network

Zhengyi Liu*

School of Computer Science and
Technology, Anhui University
Hefei, China
liuzywen@ahu.edu.cn

Yuan Wang

School of Computer Science and
Technology, Anhui University
Hefei, China
wangyuan.ahu@qq.com

Zhengzheng Tu

School of Computer Science and
Technology, Anhui University
Hefei, China
15352718@qq.com

Yun Xiao

School of Computer Science and
Technology, Anhui University
Hefei, China
280240406@qq.com

Bin Tang

School of Artificial Intelligence and
Big Data, Hefei University
Hefei, China
424539820@qq.com

ABSTRACT

Salient object detection is the pixel-level dense prediction task which can highlight the prominent object in the scene. Recently U-Net framework is widely used, and continuous convolution and pooling operations generate multi-level features which are complementary with each other. In view of the more contribution of high-level features for the performance, we propose a triplet transformer embedding module to enhance them by learning long-range dependencies across layers. It is the first to use three transformer encoders with shared weights to enhance multi-level features. By further designing scale adjustment module to process the input, devising three-stream decoder to process the output and attaching depth features to color features for the multi-modal fusion, the proposed triplet transformer embedding network (TriTransNet) achieves the state-of-the-art performance in RGB-D salient object detection, and pushes the performance to a new level. Experimental results demonstrate the effectiveness of the proposed modules and the competition of TriTransNet.¹

CCS CONCEPTS

- Computing methodologies → Interest point and salient region detections.

KEYWORDS

salient object detection; RGB-D image; transformer; shared weights; self-attention

*Corresponding author.

¹The code is available at <https://github.com/liuzywen/TriTransNet>.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '21, October 20–24, 2021, Virtual Event, China

© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8651-7/21/10...\$15.00
<https://doi.org/10.1145/3474085.3475601>

ACM Reference Format:

Zhengyi Liu, Yuan Wang, Zhengzheng Tu, Yun Xiao, and Bin Tang. 2021. TriTransNet: RGB-D Salient Object Detection with a Triplet Transformer Embedding Network. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3474085.3475601>

1 INTRODUCTION

Salient object detection (SOD) simulates the visual attention mechanism to capture the prominent object in the scene. It has been widely applied in the computer vision tasks, such as image segmentation [18], tracking [30, 47, 83], retrieval [25], compression [32], edit [65] and quality assessment [34].

As a pixel-level dense prediction task, salient object detection usually uses CNN based U-Net framework[58] (Fig. 1(a)) to encode images from low-level to high-level, and then decode back to the full spatial resolution. Research[74] points out that the performance tends to saturate quickly when gradually aggregating features from high-level to low-level. In other words, high-level features contribute more to the performance. Therefore, we propose a triplet transformer embedding module (TTEM) to enhance the feature representation of high three layers.

As we all known, Transformer[62] has recently attracted a lot of attention in computer vision domain, but it is also encountering high computational cost problem. PVT[66] adopts a spatial-reduction attention (SRA) layer to reduce the resource cost to learn high-resolution feature maps. CvT[72] introduces convolutional into the Vision Transformer architecture to concurrently maintain a high degree of computational and memory efficiency. Swin Transformer[44] uses the shifted windows calculation method to propose a hierarchical Transformer, which has the flexibility of modelling at various scales and has linear computational complexity relative to the image size. Multi-Scale Vision Longformer[82] proposes multi-scale coding structure, and further improves its attention mechanism to reduce the computational and memory cost.

Unlike these profound designs, we introduce Transformer into U-Net framework to enhance the features of high three layers, which can be easily integrated into existing U-Net framework for significant improvement with less cost. The features of high three

layers show the different attributions but the same in nature, which are the different aspects of the same input image. The proposed triplet transformer embedding module (TTEM) is composed of three standard transformer encoders[62] with shared weights. It is beneficial to find the common information which is hidden in the multi-level features and achieve the better fusion by learning long-range dependencies across levels.

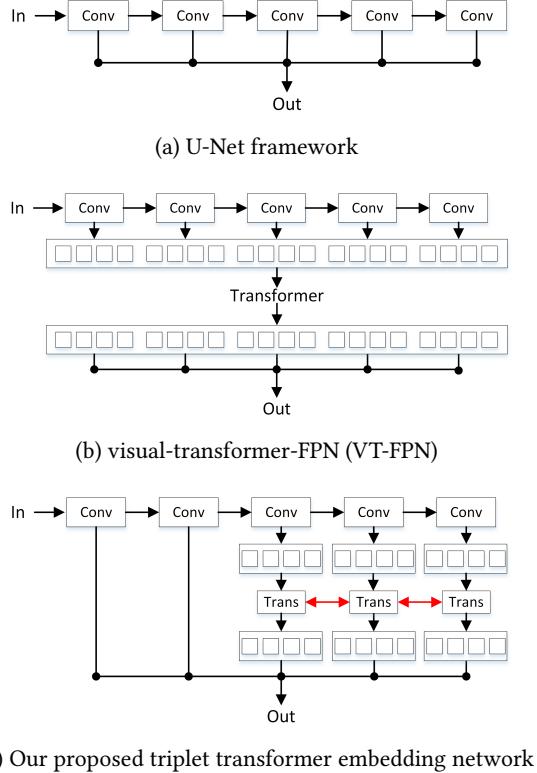


Figure 1: Comparison between U-Net framework, VT-FPN and our proposed network.

Taking TTME as the core, we further propose the triplet transformer embedding network (TriTransNet). At first, multi-level features are adjusted to the same size by a transition layer and progressively upsampling fusion module. Second, features are fed into TTME to be enhanced. Last, the output features of TTME and the features of low two layers are effectively fused by a three-stream decoder.

Our proposed TriTransNet is the first attempt to use three standard transformer encoders with shared weights to enhance the feature representation. Different from visual-transformer-FPN (VT-FPN)[71] in Fig. 1(b) which merges visual tokens from feature map of each layer with one transformer, our TriTransNet in Fig. 1(c) adopts weight sharing strategy to make visual tokens extracted from multi-level features more abundant enough to express the original information, and meanwhile high-layer semantic information and middle-layer texture or shape information are both better excavated by parallel self-attention mechanism.

In addition, depth information is proved to supply the useful cues and boost the performance for saliency detection [51], especially in some challenging and complex scenarios, e.g. the low color contrasts between salient objects and background, the cluttered background interferences. But depth image with poor quality, which likes a noise, brings some negative influences [22]. Following depth guided manners [11, 12, 23, 52, 57, 60, 79, 89, 93] we design depth purification module, which uses depth information to purify the color features.

Our main contributions can be summarized as follows:

- A triplet transformer embedding module is proposed and embedded into CNN based U-Net framework to enhance the feature representation. It is composed of three standard transformer encoders with shared weights, learning the common information from multi-level features.
- Based on the proposed triplet transformer embedding module, triplet transformer embedding network is designed to detect the salient objects in RGB-D image. Multi-level features from encoder need to be adjusted to the same size by a transition layer and progressively upsampling fusion module, and then fed into triplet transformer embedding module. Then the output of triplet transformer embedding module need to be combined with the features of low two layers by three-stream decoder to achieve the decoding process.
- Depth image is viewed as the supplement to color feature, and attached to color feature to enhance the feature representation by depth purification module which introduces spatial attention and channel attention.
- Due to the advantage of the proposed triplet transformer embedding module, the proposed model pushes the performance of RGB-D salient object detection to a new level and shows the state-of-the-art performance on several public datasets.

2 RELATED WORK

2.1 RGB-D saliency detection

In RGB-D image, color image provides appearance and texture information, and depth image contains 3D layout and spatial structure. The fusion of color feature and depth feature is always an important issue in RGB-D saliency detection. References[10, 45, 85, 90] use early fusion or input fusion, references[7, 11, 38, 39, 41] employ two-stream subnetwork to achieve the middle fusion, references[53, 57, 73, 89, 93] apply depth guided fusion and references [14, 46, 64] adopt late fusion.

Although depth information can supply the useful cues for saliency detection [51], depth image with poor quality can bring some negative influences too [22]. In order to solve the filtering issue of low-quality depth map, D3Net [22] uses gate mechanism to filter the poor depth map, EF-Net [9] enhances the depth maps by color hint map, DQSD [5] integrates a depth quality aware subnetwork into the classic bi-stream structure, assigning the weight of depth feature before conducting the fusion. In addition, CoNet[33], DASNet [88], SSDP[68] and MobileSal [73] introduce depth estimation, learning to detect the salient object simultaneously.

In the paper, we adopt depth guided manner. Depth information is viewed as the supplement to the color feature. It enhances the color feature by attention mechanism.

2.2 Transformer

Transformer is first proposed by[62] to replace recurrent neural networks (RNN), e.g.long short-term memory (LSTM) and gated recurrent unit(GRU) for machine translation tasks. It can overcome intrinsic shortages of RNN and has dominated nature language processing (NLP) field and are becoming increasingly popular in computer vision tasks, e.g. image classification[19], object detection[4], semantic segmentation[?], line segment[77], person re-identification[94], action detection[87], image completion[91], 3D point cloud processing[26, 86], pose estimation[59], facial expression recognition[48], object tracking[49] etc. DETR[4] takes the lead in applying Transformer to the field of object detection and achieves the better performance. The successful use of ViT[19] in image classification tasks has made the research on visual Transformer a hot topics. SETR[92] deploys a pure Transformer as the encoder, combined with a simple decoder, to achieve a powerful semantic segmentation model. Besides, TransUNet[8] uses the pre-trained ViT[19] as a powerful backbone of the U-Net[58] network structure, and performs well in the field of medical image segmentation.

However, pure transformer has great limitations. As a result, many improved visual transformers have emerged. The Conditional Position encodings Visual Transformer (CPVT)[17] replaces the fixed position encoding in ViT[19] with the proposed conditional position encoding (CPE), which makes it possible for Transformer to process inputs of arbitrary sizes. Tokens-to-Token (T2T)[78] adopts a novel progressive tetanization mechanism, which models local structural information by aggregating adjacent tokens into one token, while reducing the length of the token. LocalViT[42] adds locality to vision transformers by introducing depth-wise convolution into the feed-forward network, improving a locality mechanism for information exchange within a local region. Considering that most visual Transformers ignore the inherent structural information inside the sequence of patches, Transformer-iN-Transformer (TNT)[27] proposes to use outer Transformer block and inner Transformer block to model patch-level and pixel-level representations, respectively. Co-Scale Conv-Attentional Image Transformers[76] designs a conv-attention module to realize relative position embedding and enhance computation efficiency, and further proposes a co-scale mechanism to introduce cross-scale attention to enrich multi-scale feature.

On the other hand, CNN has the advantages of extracting low level features and strengthening locality, while Transformer has the advantages in establishing long-range dependencies. Some research makes full use of both advantages. TransFuse[84] uses a dual-branch structure, which uses Transformer to capture global dependencies, while low-level spatial details are extracted by CNN branches. Similarly, CoTr[75] uses the CNN backbone to extract feature representations and proposes to use deformable Transformer (DeTrans) to model long-range dependencies, effectively bridging the convolutional neural network and Transformer. ICT[63] uses transformer to recover pluralistic coherent structures together with

some coarse textures, and uses CNN to enhances local texture details of coarse priors, so as to achieve excellent results on the image completion task. TransT[13] uses Siamese-based CNN network for feature extraction, and designs the self-attention-based ego-context augment (ECA) and cross-attention-based cross-feature augment (CFA) modules for feature fusion. Compact Transformers[28] eliminates the requirement for class token and position embedding through a novel sequence pooling strategy and the use of convolutions, so as to perform head-to-head with state-of-the-art CNNs on small datasets.

Follow this strategy, we present triplet transformer embedding module which is embedded into a U-Net framework to improve the performance of RGB-D saliency detection. Combining both advantages, our model achieves the state-of-the-art performance.

3 PROPOSED METHOD

3.1 Overview

The overall framework of the proposed triplet transformer embedding network is depicted in Fig.2(a), which consists of multi-modal fusion encoder, feature enhancement module and three-stream decoder. The details can be seen in the following sections.

3.2 Multi-modal fusion encoder

Color and depth image in RGB-D image are two expressions for different modalities of the same scene. Color image provides appearance cue and depth image shows three dimension spatial information. Due to existence of poor quality depth map induced by the imaging devices or conditions, we propose multi-modal fusion encoder, in which depth features are first purified by multi-modal features using attention mechanism, and then served as supplement to the color feature by the residual connection[29]. Residual part is designed as depth purification module (DPM), and shortcut connection part is used to preserve more original color information.

In DPM which is shown in Fig. 2(b), depth feature is concatenated with color feature, and fed into a channel attention module to get attentive channel mask, which is used to purify the depth feature in a channel manner. Next, purified depth feature is fed into a spatial attention module again to generate attentive spatial mask, which is used to purify the depth feature in a spatial manner. The process can be described as:

$$F_i^r = f_i^d \times SA(f_i^d \times CA(Cat(f_i^d, f_i^r))) + f_i^r \quad (1)$$

where f_i^r and f_i^d represent color and depth features extracted by backbone network respectively in which $i = 1, \dots, 5$, $Cat(\cdot)$ denotes concatenation and following convolution operation, $CA(\cdot)$ and $SA(\cdot)$ are channel and spatial attention operation which is proposed by CBAM[70], “ \times ” is element-wise multiplication operation, “+” is element-wise addition operation.

Thus, the depth feature with poor quality can be purified, and then attached to color feature to generate more accuracy feature representation $F_i^r (i = 1, \dots, 5)$.

3.3 Feature enhancement module

In this module, we first adjust the features of high three layers to the same size, and then use the triplet transformer embedding

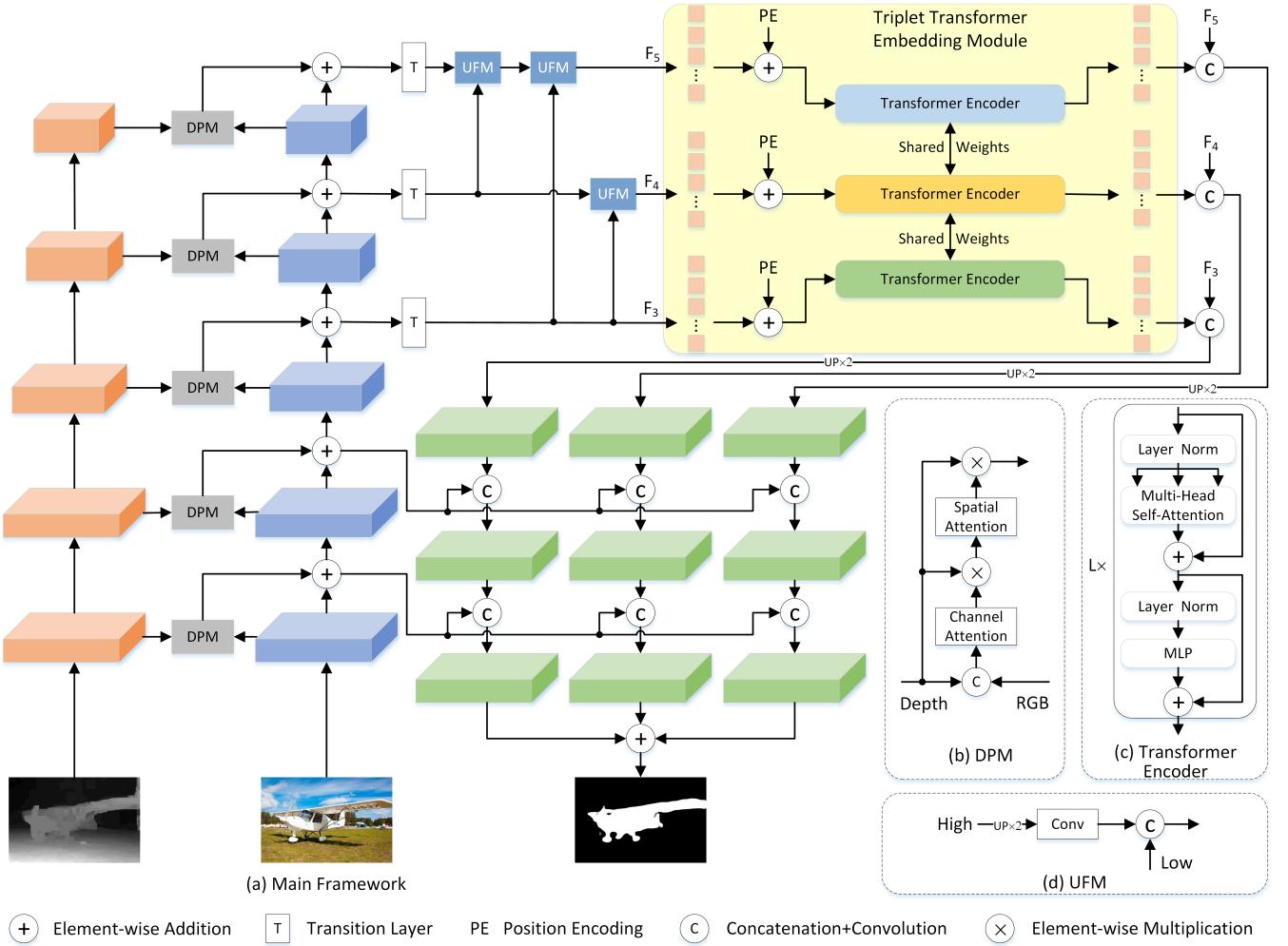


Figure 2: Our proposed triplet transformer embedding network for RGB-D salient object detection.

module to enhance the feature representation by learning long-range dependency across levels, and last concatenate the input and output of triplet transformer embedding module to preserve more original information.

3.3.1 Scale adjustment module. The triplet transformer embedding module is composed of three standard transformer encoders with shared weights. Its input should be the features with the same size. But the sizes of the multi-level features F_i^r from multi-modal fusion encoder are the different. Therefore, the first important task is to adjust the sizes of multi-level features.

At first, a transition layer which contains a 3×3 convolution and a ReLU activation function is applied on F_i^r . It can adjust the number of channels of multi-level features to the same size. It can be described as:

$$F_i'^r = \sigma(Conv(F_i^r)) \quad i = 3, \dots, 5 \quad (2)$$

where $Conv(\cdot)$ is 3×3 convolution operation, and $\sigma(\cdot)$ is ReLU activation function.

Then, we design a progressively upsampling fusion module which is used to adjust the resolution of the features in the high three layers to the same size. Since the direct upsampling with $2 \times$ or $4 \times$ ratio will bring some noises, the features are progressively upsampled and fused. The fusion process can be described as:

$$\begin{aligned} F_5 &= UFM(UFM(F_5'^r, F_4'^r), F_3'^r) \\ F_4 &= UFM(F_4'^r, F_3'^r) \\ F_3 &= F_3'^r \end{aligned} \quad (3)$$

where $UFM(\cdot)$ is shown in Fig.2(d). The detail can be described as:

$$UFM(F_{high}, F_{low}) = Cat(Conv(Up(F_{high})), F_{low}) \quad (4)$$

where F_{high} and F_{low} denote the feature from the higher layer with low resolution and the feature from the lower layer with high resolution, respectively, and $Up(\cdot)$ denotes $2 \times$ upsampling operation.

Compared with direct $2 \times, 4 \times$ upsampling on $F_4'^r$ and $F_5'^r$, progressively upsampling fusion module can not only adjust the features

to the same resolution but also increase the spatial detail of feature in the high layer by progressive fusion process.

Thus, the features $F_i (i = 3, \dots, 5)$ with the same scales will be served as the input and fed into next triplet transformer embedding module.

3.3.2 Triplet Transformer Embedding Module (TTEM). The features are first converted into the sequences of feature embedding, and then fed to three standard transformer encoders with shared weights to model the long-range relationship among different levels, and last reshaped to the original size of features.

Specifically, each input feature $F_i (i = 3, \dots, 5)$ are first flattened into a 1D sequence $\{F_i^p | p = 1, \dots, N\}$, where N is the number of patches. Each patch F_i^p is then mapped into a latent D -dimensional embedding space by a trainable linear projection layer. Furthermore, we learn specific position embedding which are added to the patch embedding to retain positional information. The process can be described as:

$$Z_i^0 = [F_i^1 + PE^1; F_i^2 + PE^2; \dots; F_i^N + PE^N] \quad (5)$$

where $PE = \{PE^p | p = 1, \dots, N\}$ is a 1D learnable positional embedding.

The remaining architecture essentially follows the standard transformer encoder[62] which stacks L transformer layer. It is shown in Fig.2(c). Each transformer layer contains multi-headed self-attention (MSA) and multi-layer perceptron (MLP) sublayer. Layer normalization (LN)[2] are inserted before these two sublayers, and the residual connection is performed after these two sublayers. The process can be described as:

$$\begin{cases} Z_i^{l'} = MSA \left(LN \left(Z_i^{l-1} \right) \right) + Z_i^{l-1} \\ Z_i^l = MLP \left(LN \left(Z_i^{l'} \right) \right) + Z_i^{l'} \end{cases} \quad l = 1, \dots, L \quad (6)$$

where L denotes the number of transformer layers in the standard transformer encoder.

3.3.3 Feature concatenation module. The outputs of three weights shared transformer encoders $Z_i^L (i = 3, \dots, 5)$ fuses the information of three layers by Transformer mechanism, so as to enhance the original feature representation. In order to preserve the more original information, we further cascade these outputs with original features to generate the enhanced features of high three layers. The process can be described as:

$$F'_i = Cat(Z_i^L, F_i) \quad i = 3, \dots, 5 \quad (7)$$

3.4 Three-stream decoder

After the features of high three layers are enhanced by the proposed triplet transformer embedding module, we will combine them with the features of low two layers to achieve the decoding process. There are two decoding methods. One is single-stream decoding and the other is three-stream decoding. The single-stream decoding first fuses three output results of feature enhancement module, and then combine it with two features in the low layers. The three-stream decoding first combines each output result of feature enhancement module with two features in the low layers, and then fuses three-stream results. We conduct two decoding processes, and find three-stream decoding is better than single-stream decoding. Next, we

use formula to show three-stream decoder as follow:

$$F''_i = Cat(Cat(Up(F'_i), F_r^2), F_r^1) \quad i = 3, \dots, 5 \quad (8)$$

The above three features are performed upsampling, convolution operation and sigmoid function to generate the saliency maps $S_i (i = 1, \dots, 3)$ which are supervised by the ground truth maps.

$$S_i = sig(Conv(Up(Conv(Up(F''_i)))))) \quad (9)$$

where $sig(\cdot)$ denotes sigmoid function.

At last, we also fuse all the features above to generate the final saliency map.

$$S_{final} = sig \left(\sum_{i=3}^5 Conv(Up(Conv(Up(F''_i)))) \right) \quad (10)$$

Pixel position aware loss L_{ppa}^s [69] is adopted for end-to-end training. The whole loss is defined as:

$$L = L_{ppa}^s(S_{final}, G) + \sum_{i=3}^5 L_{ppa}^s(S_i, G) \quad (11)$$

where G is ground truth saliency map.

4 EXPERIMENTS

4.1 Datasets and evaluation metrics

4.1.1 Datasets. We evaluate the proposed method on six challenging RGB-D SOD datasets. NLPR [54] includes 1000 images with single or multiple salient objects. NJU2K [36] consists of 2003 stereo image pairs and ground-truth maps with different objects, complex and challenging scenes. STERE [50] incorporates 1000 pairs of binocular images downloaded from the Internet. DES [15] has 135 indoor images collected by Microsoft Kinect. SIP [22] contains 1000 high-resolution images of multiple salient persons. DUT [56] contains 1200 images captured by Lytro camera in real life scenes.

For the sake of fair comparison, we use the same training dataset as in [11, 22], which consists of 1,485 images from the NJU2K dataset and 700 images from the NLPR dataset. The remaining images in the NJU2K and NLPR datasets and the whole datasets of STERE, DES and SIP are used for testing. In addition, on the DUT dataset, we follow the same protocols as in [33, 39, 56, 57, 90] to add additional 800 pairs from DUT for training and test on the remaining 400 pairs. In summary, our training set contains 2,185 paired RGB and depth images, but when testing is conducted on DUT, our training set contains 2,985 paired ones.

4.1.2 Evaluation Metrics. We adopt five widely used metrics to evaluate the performance of our model and other state-of-the-art RGB-D SOD models, including the precision-recall(PR) curve [3], E-measure [21], S-measure [20], F-measure [1] and mean absolute error (MAE) [55]. Specifically, the PR curve plots precision and recall values by setting a series of thresholds on the saliency maps to get the binary masks and further comparing them with the ground truth maps. The E-measure simultaneously captures global statistics and local pixel matching information. The S-measure can evaluate both region-aware and object-aware structural similarity between saliency map and ground truth. The F-measure is the weighted harmonic mean of precision and recall, which can evaluate the overall performance. The MAE measures the average of the per-pixel absolute difference between the saliency maps and the ground

truth maps. In our experiment, E-measure and F-measure adopts adaptive values.

4.2 Implementation details

During the training and testing phase, the input RGB and depth images are resized to 256×256 . Multiple enhancement strategies are used for all training images, i.e. random flipping, rotating and border clipping. Parameters of the backbone network are initialized with the pretrained parameters of ResNet-50 network [29]. The hyper-parameters in transformer encoder are set as: $L = 12, D = 768, N = 1024$. The rest of parameters are initialized to PyTorch default settings. We employ the Adam optimizer [37] to train our network with a batch size of 3 and an initial learning rate $1e-5$, and the learning rate will be divided by 10 every 60 epochs. Our model is trained on a machine with a single NVIDIA GTX 3090 GPU. The model converges within 150 epochs, which takes nearly 15 hours.

4.3 Comparisons with the state-of-the-art

Our model is compared with 16 state-of-the-art RGB-D SOD models, including D3Net [22], ICNet [41], DCMF [6], DRLF [67], SSF [81], SSMA [43], A2dete [57], UCNet [80], CoNet [33], DANet [90], JLDCF[24], EBFSP[31], CDNet[35], HAINet[40], RD3D[10] and DSA2F[61]. To ensure the fairness of the comparison results, the saliency maps of the evaluation are provided by the authors or generated by running source codes.

4.3.1 Quantitative Evaluation. Figure.3 shows the comparison results on PR curve. Table.1 shows the quantitative comparison results of four evaluation metrics. As can be clearly observed from figure that our curves are very short, which means that our recall is very high. Furthermore, from the table, we can see that all the evaluation metrics are nearly the best on six datasets, so as to verify the effectiveness and advantages of our proposed method. Only two S-measure values in NLPR and STERE datasets are inferior to the best, but they are also the second best. Combined with the results of figure and table, our method achieves the impressive performance.

4.3.2 Qualitative Evaluation. To make the qualitative comparisons, we show some visual examples in Figure.4. It can be observed that our method has better detection results than other methods in some challenging cases: similar foreground and background(1^{st} - 2^{nd} rows), complex scene(3^{rd} - 4^{th} rows), low quality depth map(5^{th} - 6^{th} rows), small object(7^{th} - 8^{th} rows) and multiple objects(9^{th} - 10^{th} rows). These indicate that our approach can better locate salient objects and produce more accurate saliency maps. In addition, our approach can produce more fine-grained details as highlighted in the salient region(11^{th} - 12^{th} rows). This is also the proof of the effectiveness of our method.

4.4 Ablation studies

We conduct ablation studies on NLPR, NJU2K, SIP and STERE datasets to investigate the contributions of different modules in the proposed method.

4.4.1 The effectiveness of triplet transformer embedding module (TTEM). The baseline model used here removes TTEM. Its performance is shown in the variant No.1 of Table. 2. Further, we replace

TTEM with gated recurrent unit (GRU) [16], whose result is shown in the variant No.2 of Table. 2. The variant No.3 of Table. 2 is the result of siamese transformer applied in the high two layers. The variant No.4 of Table. 2 is the result of quadruplet transformer applied in the high four layers. The variant No.5 of Table. 2 is our result of triplet transformer applied in the high three layers.

It can be clearly observed that compared with No.1, the result of our TriTransNet is improved 0.016 in the S-measure metric, 0.021 in the F-measure metric, 0.008 in the E-measure metric and 0.007 in the MAE metric on average. Meanwhile, compared with No.2, the result of our TriTransNet is improved 0.012 in the S-measure metric, 0.014 in the F-measure metric, 0.006 in the E-measure metric and 0.005 in the MAE metric on average. TTEM plays an important role in the performance improvement.

In addition, we compare No.3, No.4 and No.5 and find that Triplet win Siamese in S-measure, F-measure, E-measure, and MAE about 0.009, 0.016, 0.006 and 0.005 on average, and outperform Quadruplet about 0.010, 0.009, 0.005 and 0.004 on average. Our TriTransNet enhances long-range dependency of semantic information by using the features in the high three layers, and further combines with three-stream usampling decoding in the low two layers to perfectly depict the detailed boundary, so as to achieve the best performance.

4.4.2 The effectiveness of three-stream decoder. we further conduct the ablation study by replacing three stream decoder with single-stream decoder to check the effectiveness of the designed three-stream decoder. Table. 3 No.1 denotes the model which adopts single-stream decoder and No.2 means our three-stream decoder. From Table. 3, we can see that the use of three-stream decoder obviously improves the detection performance. It benefits from the full integration of multi-layer features.

4.4.3 The effectiveness of depth purification module (DPM). The baseline model used here removes depth purification module (DPM). It attaches the depth feature to color feature by element-wise addition operation in the encoder. Its performance is illustrated in the variant No.1 of Table. 4. Further, we discuss the similar depth-enhanced module (DEM) proposed in BBS[23] whose result is shown in the variant No.2 of Table. 4. The variant No.3 of Table. 4 denotes the model which adopts DPM instead of element-wise addition operation based on the baseline.

Compared with No.1, the performance of the variant No.3 is significantly improved. Meanwhile, compared with No.2 which using DEM, our detection effect is also better than that of No.2. It verified that the effectiveness of DPM.

5 CONCLUSIONS

In the paper, we introduce transformer into U-Net framework to detect salient object in RGB-D image. Different from existing combination method of transformer and convolutional neural networks, we propose a triplet transformer embedding module which can be embedded into existing U-Net models for the better feature representation by learning long-range dependency among different levels with less cost. Furthermore, we use depth information to enhance RGB features by depth purification module. Experimental results show our method pushes the performance to a new level,

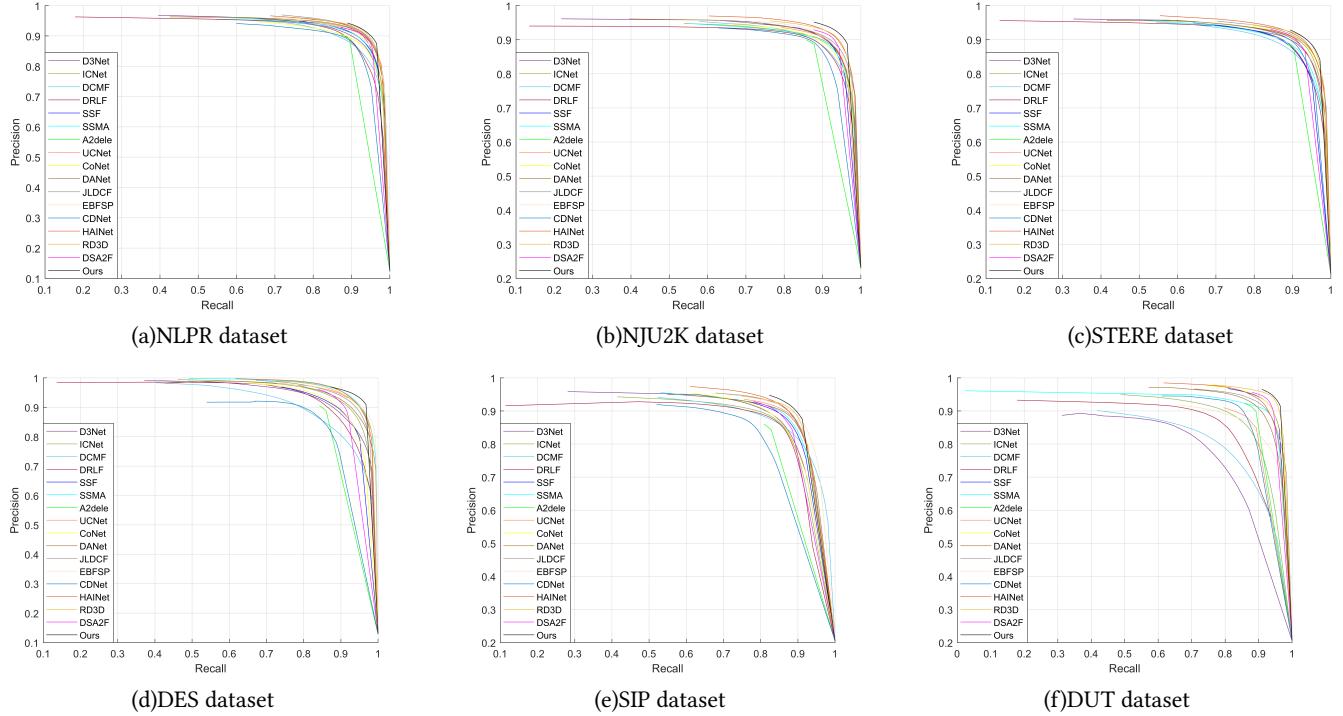


Figure 3: P-R curves comparisons of different models on six datasets.

Table 1: S-measure, adaptive F-measure, adaptive E-measure, MAE comparisons with different models. The best result is in bold.

Datasets	Metric	D3Net	ICNet	DCMF	DRFL	SSF	SSMA	A2dele	UCNet	CoNet	DANet	JLDCF	EBFSP	CDNet	HAINet	RD3D	DSA2F	TriTransNet	Ours
NLPR	S↑	.912	.923	.900	.903	.914	.915	.896	.920	.908	.920	.925	.915	.902	.924	.930	.918	.928	
	F β ↑	.861	.870	.839	.843	.875	.853	.878	.890	.846	.875	.878	.897	.848	.897	.892	.892	.909	
	E ξ ↑	.944	.944	.933	.936	.949	.938	.945	.953	.934	.951	.953	.952	.935	.957	.958	.950	.960	
	MAE↓	.030	.028	.035	.032	.026	.030	.028	.025	.031	.027	.022	.026	.032	.024	.022	.024	.020	
NJU2K	S↑	.901	.894	.889	.886	.899	.894	.869	.897	.895	.899	.902	.903	.885	.912	.916	.904	.920	
	F β ↑	.865	.868	.859	.849	.886	.865	.874	.889	.872	.871	.885	.894	.866	.900	.901	.898	.919	
	E ξ ↑	.914	.905	.897	.901	.913	.896	.897	.903	.912	.908	.913	.907	.911	.922	.918	.922	.925	
	MAE↓	.046	.052	.052	.055	.043	.053	.051	.043	.046	.045	.041	.039	.048	.038	.036	.039	.030	
STERE	S↑	.899	.903	.883	.888	.887	.890	.878	.903	.905	.901	.903	.900	.896	.907	.911	.897	.908	
	F β ↑	.859	.865	.841	.845	.867	.855	.874	.885	.884	.868	.869	.870	.873	.885	.886	.893	.893	
	E ξ ↑	.920	.915	.904	.915	.921	.907	.915	.922	.927	.921	.919	.912	.922	.925	.927	.927	.927	
	MAE↓	.046	.045	.054	.050	.046	.051	.044	.039	.037	.043	.040	.045	.042	.040	.037	.039	.033	
DES	S↑	.898	.920	.877	.895	.905	.941	.885	.933	.911	.924	.931	.937	.975	.935	.935	.916	.943	
	F β ↑	.870	.889	.820	.868	.876	.906	.865	.917	.861	.899	.900	.913	.839	.924	.917	.901	.936	
	E ξ ↑	.951	.959	.923	.954	.948	.974	.922	.974	.945	.968	.969	.974	.921	.974	.975	.955	.981	
	MAE↓	.031	.027	.040	.030	.025	.021	.028	.018	.027	.023	.020	.018	.034	.018	.019	.023	.014	
SP	S↑	.860	.854	.859	.850	.868	.872	.826	.875	.858	.875	.880	.885	.823	.880	.885	.862	.886	
	F β ↑	.835	.836	.819	.813	.851	.854	.825	.868	.842	.855	.873	.869	.805	.875	.874	.865	.892	
	E ξ ↑	.902	.899	.898	.891	.911	.911	.892	.913	.909	.914	.921	.917	.880	.919	.920	.908	.924	
	MAE↓	.063	.069	.068	.071	.056	.057	.070	.051	.063	.054	.049	.044	.076	.053	.048	.057	.043	
DUT	S↑	.775	.852	.798	.826	.916	.903	.886	.919	.899	.906	.858	.880	.910	.931	.921	.933		
	F β ↑	.756	.830	.750	.803	.914	.866	.890	.856	.909	.888	.882	.842	.874	.906	.924	.926	.938	
	E ξ ↑	.847	.897	.848	.870	.946	.921	.924	.903	.948	.934	.931	.890	.918	.938	.949	.950	.957	
	MAE↓	.097	.072	.104	.080	.034	.044	.043	.056	.033	.043	.067	.067	.048	.038	.031	.030	.025	

and ablation studies also verify the effectiveness of each module. In the future, we will achieve the same task by a pure transformer, and further discuss their respective advantages to achieve the better combination.

ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China (62006002), Natural Science Foundation of Anhui Province (1908085MF182) and Key Program of Natural Science Project of Educational Commission of Anhui Province(KJ2019A0034).

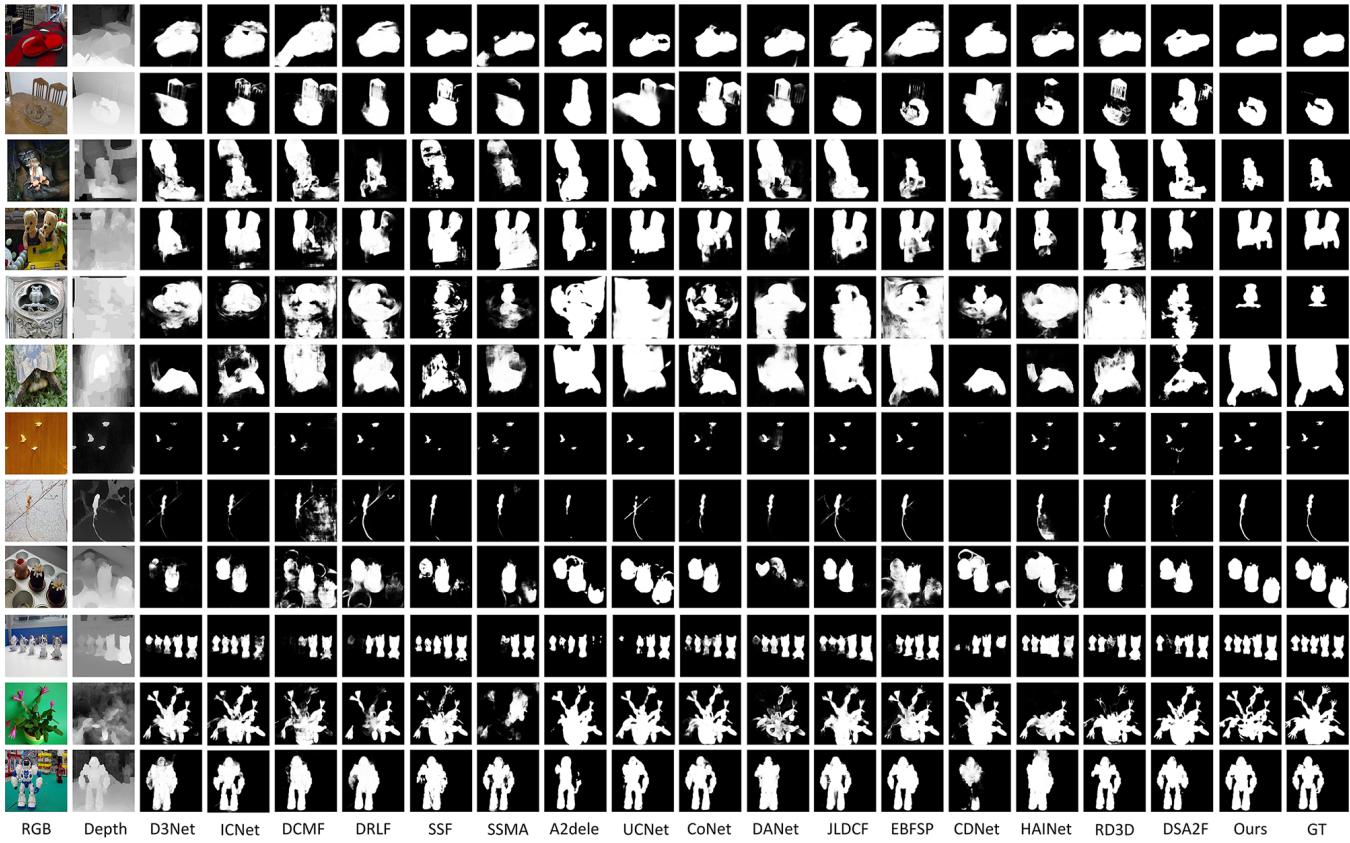


Figure 4: Visual comparison results with other the state-of-the-art models.

Table 2: Ablation experiment of triplet transformer embedding module (TTEM). The best result is in bold.

Variant	Candidate	NLPR			NJUD2K			SIP			STERE											
		Baseline	GRU	Siamese	Quadruplet	Triplet	S \uparrow	F β \uparrow	E ξ \uparrow	MAE \downarrow	S \uparrow	F β \uparrow	E ξ \uparrow	MAE \downarrow	S \uparrow	F β \uparrow	E ξ \uparrow	MAE \downarrow				
No.1	✓						.910	.882	.952	.026	.904	.897	.917	.038	.876	.877	.918	.049	.888	.873	.916	.042
No.2	✓	✓					.914	.891	.953	.024	.905	.901	.919	.037	.879	.882	.919	.047	.895	.883	.920	.038
No.3	✓		✓				.917	.888	.956	.024	.910	.903	.915	.035	.882	.885	.926	.046	.896	.872	.917	.040
No.4	✓			✓			.922	.903	.958	.022	.911	.908	.922	.034	.875	.886	.913	.048	.895	.881	.922	.038
No.5	✓				✓		.928	.909	.960	.020	.920	.919	.925	.030	.886	.892	.924	.043	.908	.893	.927	.033

Table 3: Ablation experiment of three-stream decoder. The best result is in bold.

Variant	Candidate	NLPR			NJUD2K			SIP			STERE		
	single-stream three-stream	$S\uparrow$	$F_B\uparrow$	$E_{\xi}\uparrow$	$MAE\downarrow$	$S\uparrow$	$F_B\uparrow$	$E_{\xi}\uparrow$	$MAE\downarrow$	$S\uparrow$	$F_B\uparrow$	$E_{\xi}\uparrow$	$MAE\downarrow$
No.1	✓	.923	.892	.958	.022	.916	.904	.919	.035	.884	.886	.920	.045
No.2	✓	.928	.909	.960	.020	.920	.919	.925	.030	.886	.892	.924	.043

Table 4: Ablation experiment of depth purification module (DPM). The best result is in bold.

Variant	Candidate			NLPR			NJUD2K			SIP			STERE						
	Baseline	DEM	DPM	S↑	F _B ↑	E _X ↑	MAE↓	S↑	F _B ↑	E _X ↑	MAE↓	S↑	F _B ↑	E _X ↑	MAE↓				
No.1	✓			.917	.897	.956	.023	.909	.904	.920	.035	.883	.887	.921	.044	.894	.875	.918	.039
No.2	✓	✓		.923	.901	.958	.021	.914	.910	.922	.033	.884	.889	.923	.044	.905	.889	.925	.035
No.3	✓		✓	.928	.909	.960	.020	.920	.919	.925	.030	.886	.892	.924	.043	.908	.893	.927	.033

REFERENCES

- [1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. 2009. Frequency-tuned salient region detection. In *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 1597–1604.
- [2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).
- [3] Ali Borji, Ming-Ming Cheng, Huaiyu Jiang, and Jia Li. 2015. Salient object detection: A benchmark. *IEEE transactions on image processing* 24, 12 (2015), 5706–5722.
- [4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision*. Springer, 213–229.
- [5] Chengzhao Chen, Jipeng Wei, Chong Peng, and Hong Qin. 2021. Depth-Quality-Aware Salient Object Detection. *IEEE Transactions on Image Processing* 30 (2021), 2350–2363.
- [6] Hao Chen, Yongjian Deng, Youfu Li, Tzu-Yi Hung, and Guosheng Lin. 2020. RGBD salient object detection via disentangled cross-modal fusion. *IEEE Transactions on Image Processing* 29 (2020), 8407–8416.
- [7] Hao Chen and Youfu Li. 2018. Progressively complementarity-aware fusion network for RGB-D salient object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3051–3060.
- [8] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. 2021. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306* (2021).
- [9] Qian Chen, Keren Fu, Ze Liu, Geng Chen, Hongwei Du, Bensheng Qiu, and Ling Shao. 2020. EF-Net: A novel enhancement and fusion network for RGB-D saliency detection. *Pattern Recognition* (2020), 107740.
- [10] Qian Chen, Ze Liu, Yi Zhang, Keren Fu, Qijun Zhao, and Hongwei Du. 2021. RGB-D Salient Object Detection via 3D Convolutional Neural. *AAAI* (2021).
- [11] Shuhuan Chen and Yun Fu. 2020. Progressively guided alternate refinement network for RGB-D salient object detection. In *European Conference on Computer Vision*. Springer, 520–538.
- [12] Sihan Chen, Xinxin Zhu, Wei Liu, Xingjian He, and Jing Liu. 2021. Global-Local Propagation Network for RGB-D Semantic Segmentation. *arXiv preprint arXiv:2101.10801* (2021).
- [13] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. 2021. Transformer tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8126–8135.
- [14] Zuyao Chen, Runmin Cong, Qianqian Xu, and Qingming Huang. 2020. DPANet: Depth Potentially-Aware Gated Attention Network for RGB-D Salient Object Detection. *IEEE Transactions on Image Processing* (2020).
- [15] Yupeng Cheng, Huazhu Fu, Xingxing Wei, Jiangjian Xiao, and Xiaochun Cao. 2014. Depth enhanced saliency detection method. In *Proceedings of international conference on internet multimedia computing and service*. 23–27.
- [16] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).
- [17] Xiangxiang Chu, Bo Zhang, Zhi Tian, Xiaolin Wei, and Huaxia Xia. 2021. Do We Really Need Explicit Position Encodings for Vision Transformers? *arXiv preprint arXiv:2102.10882* (2021).
- [18] Michael Donoser, Martin Urschler, Martin Hirzer, and Horst Bischof. 2009. Saliency driven total variation segmentation. In *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 817–824.
- [19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.
- [20] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. 2017. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision*. 4548–4557.
- [21] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. 2018. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421* (2018).
- [22] Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. 2020. Rethinking RGB-D Salient Object Detection: Models, Data Sets, and Large-Scale Benchmarks. *IEEE Transactions on Neural Networks and Learning Systems* (2020).
- [23] Deng-Ping Fan, Yingjie Zhai, Ali Borji, Jufeng Yang, and Ling Shao. 2020. BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network. In *European Conference on Computer Vision*. Springer, 275–292.
- [24] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, and Qijun Zhao. 2020. JL-DCF: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3052–3062.
- [25] Yuan Gao, Miaojing Shi, Dacheng Tao, and Chao Xu. 2015. Database saliency for fast image retrieval. *IEEE Transactions on Multimedia* 17, 3 (2015), 359–369.
- [26] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. 2020. PCT: Point Cloud Transformer. *arXiv preprint arXiv:2012.09688* (2020).
- [27] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. 2021. Transformer in transformer. *arXiv preprint arXiv:2103.00112* (2021).
- [28] Ali Hassani, Steven Walton, Nikhil Shah, Abulkemu Abuduweili, Jiachen Li, and Humphrey Shi. 2021. Escaping the Big Data Paradigm with Compact Transformers. *arXiv preprint arXiv:2104.05704* (2021).
- [29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [30] Seunghoon Hong, Tackgeun You, Suha Kwak, and Bohyung Han. 2015. Online tracking by learning discriminative saliency map with convolutional neural network. In *International conference on machine learning*. 597–606.
- [31] Nianchang Huang, Yang Yang, Dingwen Zhang, Qiang Zhang, and Jungong Han. 2021. Employing Bilinear Fusion and Saliency Prior Information for RGB-D Salient Object Detection. *IEEE Transactions on Multimedia* (2021).
- [32] Qing-Ge Ji, Zhi-Dang Fang, Zhen-Hua Xie, and Zhe-Ming Lu. 2013. Video abstraction based on the visual attention model and online clustering. *Signal Processing: Image Communication* 28, 3 (2013), 241–253.
- [33] Wei Ji, Jingjing Li, Miao Zhang, Yongqi Piao, and Huchuan Lu. 2020. Accurate rgb-d salient object detection via collaborative learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII* 16. Springer, 52–69.
- [34] Qiuping Jiang, Feng Shao, Weisi Lin, Ke Gu, Gangyi Jiang, and Huifang Sun. 2017. Optimizing multistage discriminative dictionaries for blind image quality assessment. *IEEE Transactions on Multimedia* 20, 8 (2017), 2035–2048.
- [35] Wen-Da Jin, Jun Xu, Qi Han, Yi Zhang, and Ming-Ming Cheng. 2021. CDNet: Complementary Depth Network for RGB-D Salient Object Detection. *IEEE Transactions on Image Processing* 30 (2021), 3376–3390.
- [36] Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu. 2014. Depth saliency based on anisotropic center-surround difference. In *2014 IEEE international conference on image processing (ICIP)*. IEEE, 1115–1119.
- [37] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [38] Chongyi Li, Runmin Cong, Sam Kwong, Junhui Hou, Huazhu Fu, Guopu Zhu, Dingwen Zhang, and Qingming Huang. 2020. ASIF-Net: Attention steered inter-weave fusion network for RGB-D salient object detection. *IEEE Transactions on Cybernetics* (2020).
- [39] Chongyi Li, Runmin Cong, Yongqi Piao, Qianqian Xu, and Chen Change Loy. 2020. RGB-D salient object detection with cross-modality modulation and selection. In *European Conference on Computer Vision*. Springer, 225–241.
- [40] Gongyang Li, Zhi Liu, Minyu Chen, Zhen Bai, Weisi Lin, and Haibin Ling. 2021. Hierarchical Alternate Interaction Network for RGB-D Salient Object Detection. *IEEE Transactions on Image Processing* 30 (2021), 3528–3542.
- [41] Gongyang Li, Zhi Liu, and Haibin Ling. 2020. ICNet: Information Conversion Network for RGB-D Based Salient Object Detection. *IEEE Transactions on Image Processing* 29 (2020), 4873–4884.
- [42] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool. 2021. LocalViT: Bringing Locality to Vision Transformers. *arXiv preprint arXiv:2104.05707* (2021).
- [43] Nian Liu, Ni Zhang, and Junwei Han. 2020. Learning Selective Self-Mutual Attention for RGB-D Saliency Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13756–13765.
- [44] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030* (2021).
- [45] Zhengyi Liu, Song Shi, Quntao Duan, Wei Zhang, and Peng Zhao. 2019. Salient object detection for RGB-D image by single stream recurrent convolution neural network. *Neurocomputing* 363 (2019), 46–57.
- [46] Zhengyi Liu, Wei Zhang, and Peng Zhao. 2020. A cross-modal adaptive gated fusion generative adversarial network for RGB-D salient object detection. *Neurocomputing* 387 (2020), 210–220.
- [47] Cong Ma, Zhenjiang Miao, Xiao-Ping Zhang, and Min Li. 2017. A saliency prior context model for real-time object tracking. *IEEE Transactions on Multimedia* 19, 11 (2017), 2415–2424.
- [48] Fuyan Ma, Bin Sun, and Shutao Li. 2021. Robust Facial Expression Recognition with Convolutional Visual Transformers. *arXiv preprint arXiv:2103.16854* (2021).
- [49] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. 2021. TrackFormer: Multi-Object Tracking with Transformers. *arXiv preprint arXiv:2101.02702* (2021).
- [50] Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. 2012. Leveraging stereopsis for saliency analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 454–461.
- [51] Nabil Ouerhani and Heinz Hugli. 2000. Computing visual attention from scene depth. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, Vol. 1. IEEE, 375–378.

- [52] Liang Pan, Xiaofei Zhou, Ran Shi, Jiyong Zhang, and Chenggang Yan. 2020. Cross-modal feature extraction and integration based RGBD saliency detection. *Image and Vision Computing* 101 (2020), 103964.
- [53] Youwei Pang, Lihe Zhang, Xiaoqi Zhao, and Huchuan Lu. 2020. Hierarchical dynamic filtering network for rgb-d salient object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*. Springer, 235–252.
- [54] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. 2014. Rgbd salient object detection: a benchmark and algorithms. In *European conference on computer vision*. Springer, 92–109.
- [55] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. 2012. Saliency filters: Contrast based filtering for salient region detection. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 733–740.
- [56] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. 2019. Depth-induced multi-scale recurrent attention network for saliency detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 7254–7263.
- [57] Yongri Piao, Zhengkun Rong, Miao Zhang, Weisong Ren, and Huchuan Lu. 2020. A2dele: Adaptive and Attentive Depth Distiller for Efficient RGB-D Salient Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9060–9069.
- [58] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.
- [59] Lucas Stoffl, Maxime Vidal, and Alexander Mathis. 2021. End-to-End Trainable Multi-Instance Pose Estimation with Transformers. *arXiv preprint arXiv:2103.12115* (2021).
- [60] Lei Sun, Kailun Yang, Xinxin Hu, Weijian Hu, and Kaiwei Wang. 2020. Real-time fusion network for RGB-D semantic segmentation incorporating unexpected obstacle detection for road-driving images. *IEEE Robotics and Automation Letters* 5, 4 (2020), 5558–5565.
- [61] Peng Sun, Wenhui Zhang, Huanyu Wang, Songyuan Li, and Xi Li. 2021. Deep RGB-D Saliency Detection with Depth-Sensitive Attention and Automatic Multi-Modal Fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1407–1417.
- [62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc., 5998–6008.
- [63] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. 2021. High-Fidelity Pluralistic Image Completion with Transformers. *arXiv preprint arXiv:2103.14031* (2021).
- [64] Ningning Wang and Xiaojin Gong. 2019. Adaptive fusion for RGB-D salient object detection. *IEEE Access* 7 (2019), 55277–55284.
- [65] Wenguan Wang, Jianbing Shen, and Haibin Ling. 2018. A deep network solution for attention and aesthetics aware photo cropping. *IEEE transactions on pattern analysis and machine intelligence* 41, 7 (2018), 1531–1544.
- [66] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122* (2021).
- [67] Xuehao Wang, Shuai Li, Chenglizhao Chen, Yuming Fang, Aimin Hao, and Hong Qin. 2020. Data-level recombination and lightweight fusion scheme for RGB-D salient object detection. *IEEE Transactions on Image Processing* 30 (2020), 458–471.
- [68] Yue Wang, Yuke Li, James H Elder, Runmin Wu, Huchuan Lu, and Lu Zhang. 2020. Synergistic saliency and depth prediction for RGB-D saliency detection. In *Proceedings of the Asian Conference on Computer Vision*. 1–17.
- [69] Jun Wei, Shuhui Wang, and Qingming Huang. 2020. F³Net: Fusion, Feedback and Focus for Salient Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 12321–12328.
- [70] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. CBAM: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*. 3–19.
- [71] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Masayoshi Tomizuka, Kurt Keutzer, and Peter Vajda. 2020. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677* (2020).
- [72] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. 2021. CvT: Introducing Convolutions to Vision Transformers. *arXiv preprint arXiv:2103.15808* (2021).
- [73] Yu-Huan Wu, Yun Liu, Jun Xu, Jia-Wang Bian, Yuchao Gu, and Ming-Ming Cheng. 2020. MobileSal: Extremely Efficient RGB-D Salient Object Detection. *arXiv preprint arXiv:2012.13095* (2020).
- [74] Zhe Wu, Li Su, and Qingming Huang. 2019. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3907–3916.
- [75] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. 2021. CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation. *arXiv preprint arXiv:2103.03024* (2021).
- [76] Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. 2021. Co-Scale Conv-Attentional Image Transformers. *arXiv preprint arXiv:2104.06399* (2021).
- [77] Yifan Xu, Weijian Xu, David Cheung, and Zhuowen Tu. 2021. Line segment detection using transformers without edges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4257–4266.
- [78] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Francis EH Tay, Jia Shi Feng, and Shuicheng Yan. 2021. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986* (2021).
- [79] Jin Zeng, Yanfeng Tong, Yunmu Huang, Qiong Yan, Wenxiu Sun, Jing Chen, and Yongtian Wang. 2019. Deep surface normal estimation with hierarchical rgb-d fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6153–6162.
- [80] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Sadat Saleh, Tong Zhang, and Nick Barnes. 2020. UC-Net: uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8582–8591.
- [81] Miao Zhang, Weisong Ren, Yongri Piao, Zhengkun Rong, and Huchuan Lu. 2020. Select, Supplement and Focus for RGB-D Saliency Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3472–3481.
- [82] Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng Gao. 2021. Multi-Scale Vision Longformer: A New Vision Transformer for High-Resolution Image Encoding. *arXiv preprint arXiv:2103.15358* (2021).
- [83] Pingping Zhang, Wei Liu, Dong Wang, Yinjie Lei, Hongyu Wang, and Huchuan Lu. 2020. Non-rigid object tracking via deep multi-scale spatial-temporal discriminative saliency maps. *Pattern Recognition* 100 (2020), 107130.
- [84] Yundong Zhang, Huiye Liu, and Qiang Hu. 2021. Transfuse: Fusing transformers and cnns for medical image segmentation. *arXiv preprint arXiv:2102.08005* (2021).
- [85] Zhao Zhang, Zheng Lin, Jun Xu, Wen-Da Jin, Shao-Ping Lu, and Deng-Ping Fan. 2021. Bilateral attention network for RGB-D salient object detection. *IEEE Transactions on Image Processing* 30 (2021), 1949–1961.
- [86] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip Torr, and Vladlen Koltun. 2020. Point transformer. *arXiv preprint arXiv:2012.09164* (2020).
- [87] Jiaojiao Zhao, Xinyu Li, Chunhui Liu, Shuai Bing, Hao Chen, Cees GM Snoek, and Joseph Tighe. 2021. TubeR: Tube-Transformer for Action Detection. *arXiv preprint arXiv:2104.00969* (2021).
- [88] Jiawei Zhao, Yifan Zhao, Jia Li, and Xiaowu Chen. 2020. Is depth really necessary for salient object detection?. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1745–1754.
- [89] Jia-Xing Zhao, Yang Cao, Deng-Ping Fan, Ming-Ming Cheng, Xuan-Yi Li, and Le Zhang. 2019. Contrast prior and fluid pyramid integration for RGBD salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3927–3936.
- [90] Xiaoqi Zhao, Lihe Zhang, Youwei Pang, Huchuan Lu, and Lei Zhang. 2020. A single stream network for robust and real-time rgb-d salient object detection. In *European Conference on Computer Vision*. Springer, 646–662.
- [91] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. 2021. TFill: Image Completion via a Transformer-Based Architecture. *arXiv preprint arXiv:2104.00845* (2021).
- [92] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6881–6890.
- [93] Chunbiao Zhu, Xing Cai, Kan Huang, Thomas H Li, and Ge Li. 2019. PDNet: Prior-model guided depth-enhanced network for salient object detection. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 199–204.
- [94] Kuan Zhu, Haiyun Guo, Shiliang Zhang, Yaowei Wang, Gaopan Huang, Honglin Qiao, Jing Liu, Jinqiao Wang, and Ming Tang. 2021. AAformer: Auto-Aligned Transformer for Person Re-Identification. *arXiv preprint arXiv:2104.00921* (2021).