# Sclera-TransFuse: Fusing Swin Transformer and CNN for Accurate Sclera Segmentation

Haiqing Li[1], Caiyong Wang[1], Guangzhe Zhao[1], Zhaofeng He[2], Yunlong Wang[3], Zhenan Sun[3]
[1]Beijing Key Laboratory of Robot Bionics and Function Research, Beijing University of Civil Engineering and Architecture, Beijing, P.R. China
[2]Beijing University of Posts and Telecommunications, Beijing, P.R. China
[3]CRIPAC, MAIS, CASIA, Beijing, P.R. China

IEEE International Joint Conference on Biometrics

## INTRODUCTION

### Challenge

1. Sclera images captured in non-constrained environments often suffer from various noise such as blur, occlusions, illumination variations, specular reflections and gaze deviation

2. Due to the limited range of receptive fields, CNNs are difficult to effectively model global semantic relevance and thus robustly resist noise interference
3. Lack of diversity in sclera datasets

### Goal

Enjoy the benefits of both CNNs and vision transformer in the feature extraction to achieve more accurate and complete sclera segmentation.

### Method

1. The encoder integrates CNNs and Transformer in parallel to simultaneously extract local detail features and model the long-range spatial dependence

2. A Cross-Domain Fusion (CDF) module is designed to efficiently fuse multi-scale features from CNNs and Transformer through information interaction and self-attention mechanism

3. Deep supervision strategies are employed to learn intermediate feature representations better and faster

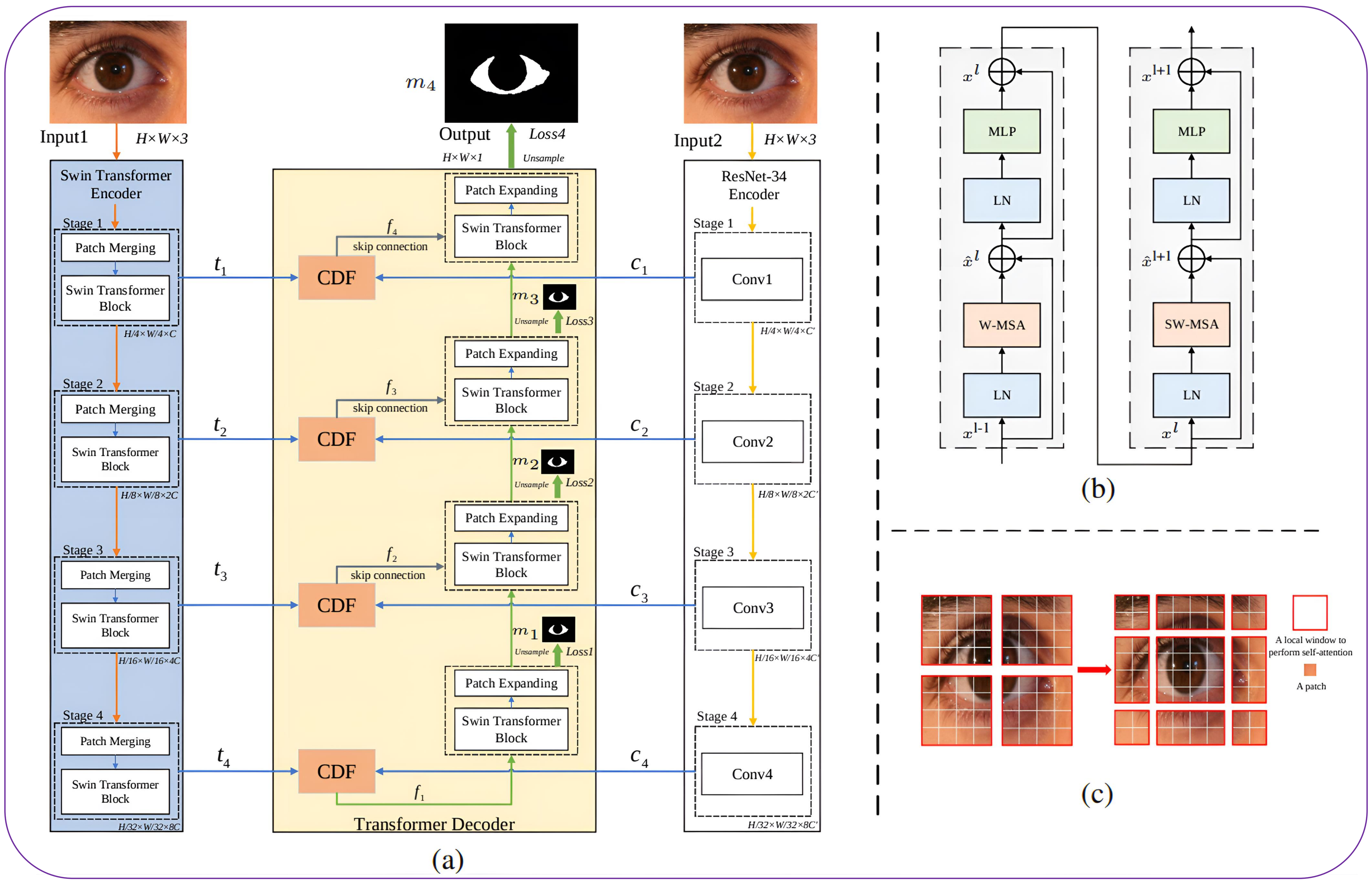
### Output

1. A novel two-stream encoder-decoder model and A novel Cross-Domain Fusion (CDF) module.

2. New performance of intra-dataset and cross-dataset evaluation settings on six common sclera segmentation datasets.

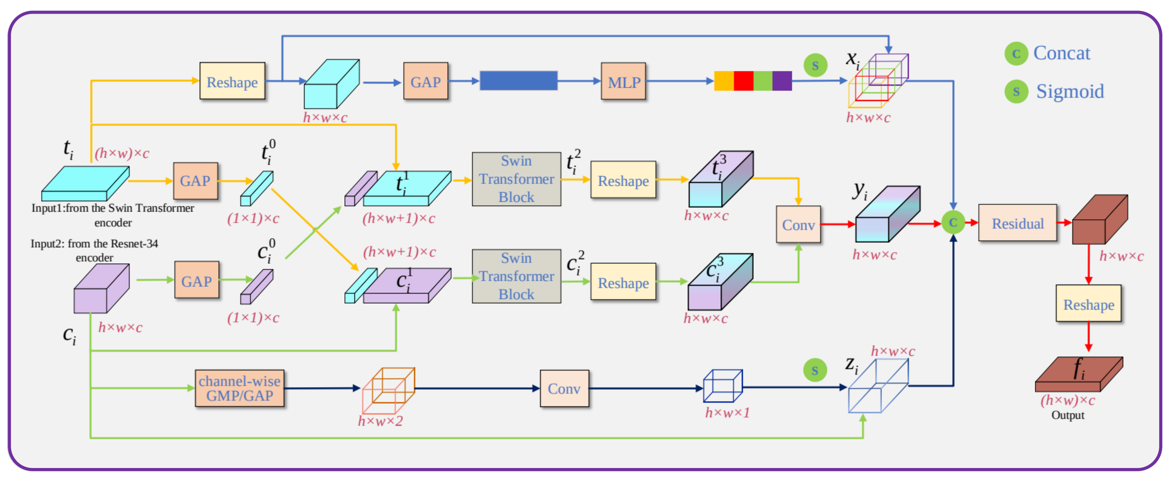
3. Greater robustness and accuracy than previous methods

## THE PROPOSED METHOD




The overall architecture of Sclera-TransFuse model

### Cross-Domain Fusion (CDF) module




The overall architecture of CDF model

**The details of CDF model**

$$t_i^0 = \mathrm{GAP}(t_i), c_i^0 = \mathrm{GAP}(c_i),$$
$$t_i^1 = \mathrm{Concat}(t_i, c_i^0), c_i^1 = \mathrm{Concat}(c_i, t_i^0),$$
$$t_i^2 = \mathrm{Swin}(t_i^1), c_i^2 = \mathrm{Swin}(c_i^1),$$
$$t_i^3 = \mathrm{Reshape}(t_i^2), c_i^3 = \mathrm{Reshape}(c_i^2),$$
$$y_i = \mathrm{Conv}(\mathrm{Concat}(t_i^3, c_i^3)),$$
$$x_i = \mathrm{Sigmoid}(\mathrm{MLP}(\mathrm{GAP}(t_i))) \otimes t_i,$$
$$z_i = \mathrm{Sigmoid}(\mathrm{Conv}(\mathrm{GAP}_c(c_i) \oplus \mathrm{GMP}_c(c_i))) \otimes c_i,$$

### Deep supervision

$$\begin{aligned}\mathcal{L}_{overall} &= \lambda_1\mathcal{L}_1 + \lambda_2\mathcal{L}_2 + \lambda_3\mathcal{L}_3 + \lambda_4\mathcal{L}_4 \\ &= \lambda_1 l(G, m_1) + \lambda_2 l(G, m_2) \\ &+ \lambda_3 l(G, m_3) + \lambda_4 l(G, m_4),\end{aligned}$$

$$l(G, m) = \mathcal{L}_{IoU}(G, m) + \mathcal{L}_{bce}(G, m),$$

## EXPERIMENTS

### Comparisons to State-Of-The-Arts

| Training dataset | Testing dataset | Method | P(%) ↑ | R(%) ↑ | F1(%) ↑ | IOU(%) ↑ |
|---|---|---|---|---|---|---|
| UBIRIS.v2 | UBIRIS.v2 | U-Net [20] | 91.53(04.96) | 90.48(06.58) | 90.95(03.16) | 83.12(07.90) |
| | | nn-UNet [13] | 91.78(02.66) | 91.23(04.75) | 91.43(02.99) | 85.96(02.40) |
| | | ScleraSegNet(SSBC) [25] | 91.94(07.27) | 91.22(06.86) | 91.28(05.56) | 84.33(07.38) |
| | | ScleraSegNet(CBAM) [25] | 91.74(07.44) | 91.24(07.26) | 91.20(05.76) | 84.21(07.63) |
| | | Sclera-TransFuse | **94.45(03.52)** | **94.47(02.60)** | **94.53(01.97)** | **89.69(02.45)** |
| MICHE-I | MICHE-I | U-Net [20] | 90.60(05.98) | 86.05(09.67) | 87.83(06.56) | 78.85(09.30) |
| | | Sclera-Net [18] | 91.88(04.23) | 94.69(04.76) | 93.13(03.93) | - |
| | | nn-UNet [13] | 90.87(03.44) | 89.05(04.80) | 90.41(04.96) | 85.22(03.64) |
| | | ScleraSegNet(SSBC) [25] | 89.31(06.12) | 90.69(07.34) | 89.69(04.88) | 81.63(07.49) |
| | | ScleraSegNet(CBAM) [25] | 91.71(05.42) | 88.11(08.26) | 89.54(05.37) | 81.45(08.03) |
| | | Sclera-TransFuse | **92.11(03.95)** | **95.69(01.94)** | **93.80(02.22)** | **88.41(03.81)** |
| SBVPI | SBVPI | U-Net [20] | 95.66(02.54) | 95.18(04.82) | 95.32(02.71) | 91.18(04.63) |
| | | Sclera-Net [18] | 94.40(03.28) | **98.17(01.60)** | 96.24(01.71) | - |
| | | nn-UNet [13] | **96.87(02.96)** | 95.12(04.31) | 95.67(03.39) | **93.88(02.97)** |
| | | ScleraSegNet(SSBC) [25] | 95.39(02.70) | 95.86(04.52) | 95.53(02.57) | 91.55(04.42) |
| | | ScleraSegNet(CBAM) [25] | 95.62(02.46) | 95.39(04.83) | 95.41(02.68) | 91.33(04.58) |
| | | Sclera-TransFuse | 96.59(01.98) | 96.82(03.33) | **96.66(02.00)** | 93.59(02.85) |
| MASD+SMD | MOBIUS | U-Net [20] | 95.05(04.69) | 70.64(06.87) | 80.48(04.30) | 77.34(04.66) |
| | | nn-UNet [13] | **95.27(03.44)** | 73.89(06.10) | 82.13(04.97) | 78.55(05.32) |
| | | ScleraSegNet(SSBC) [25] | 93.25(06.94) | 73.80(05.33) | 82.39(07.04) | 81.23(06.63) |
| | | UNet-P [22] | 90.90(04.00) | 83.10(03.00) | 86.80(03.00) | **86.80(03.00)** |
| | | Sclera-TransFuse | 89.07(10.04) | **86.23(06.81)** | **87.79(07.08)** | 85.51(10.52) |

− indicates that the value is not available in literature.

### Ablation Studies

| Method | P(%) ↑ | R(%) ↑ | F1(%) ↑ | IOU(%) ↑ |
|---|---|---|---|---|
| Sclera-TransFuse | 94.45(03.52) | **94.47(02.60)** | **94.53(01.97)** | **89.69(02.45)** |
| −ResNet-34 (encoder) | 92.74(04.22) | 93.70(02.15) | 92.84(02.35) | 86.72(04.09) |
| −Swin Transformer (encoder) | 92.52(07.35) | 93.61(04.86) | 92.71(03.98) | 86.89(06.40) |
| △ Concat+CNN (CDF) | 92.23(04.42) | 94.10(02.50) | 93.16(02.24) | 87.64(04.66) |
| − CAB (CDF) | **94.98(04.31)** | 93.37(02.02) | 94.16(02.32) | 88.79(04.15) |
| − SAB (CDF) | 94.40(04.36) | 93.56(03.25) | 93.98(02.05) | 88.03(03.96) |
| − CAB+SAB (CDF) | 92.99(04.83) | 94.18(02.06) | 93.58(02.30) | 87.83(04.58) |

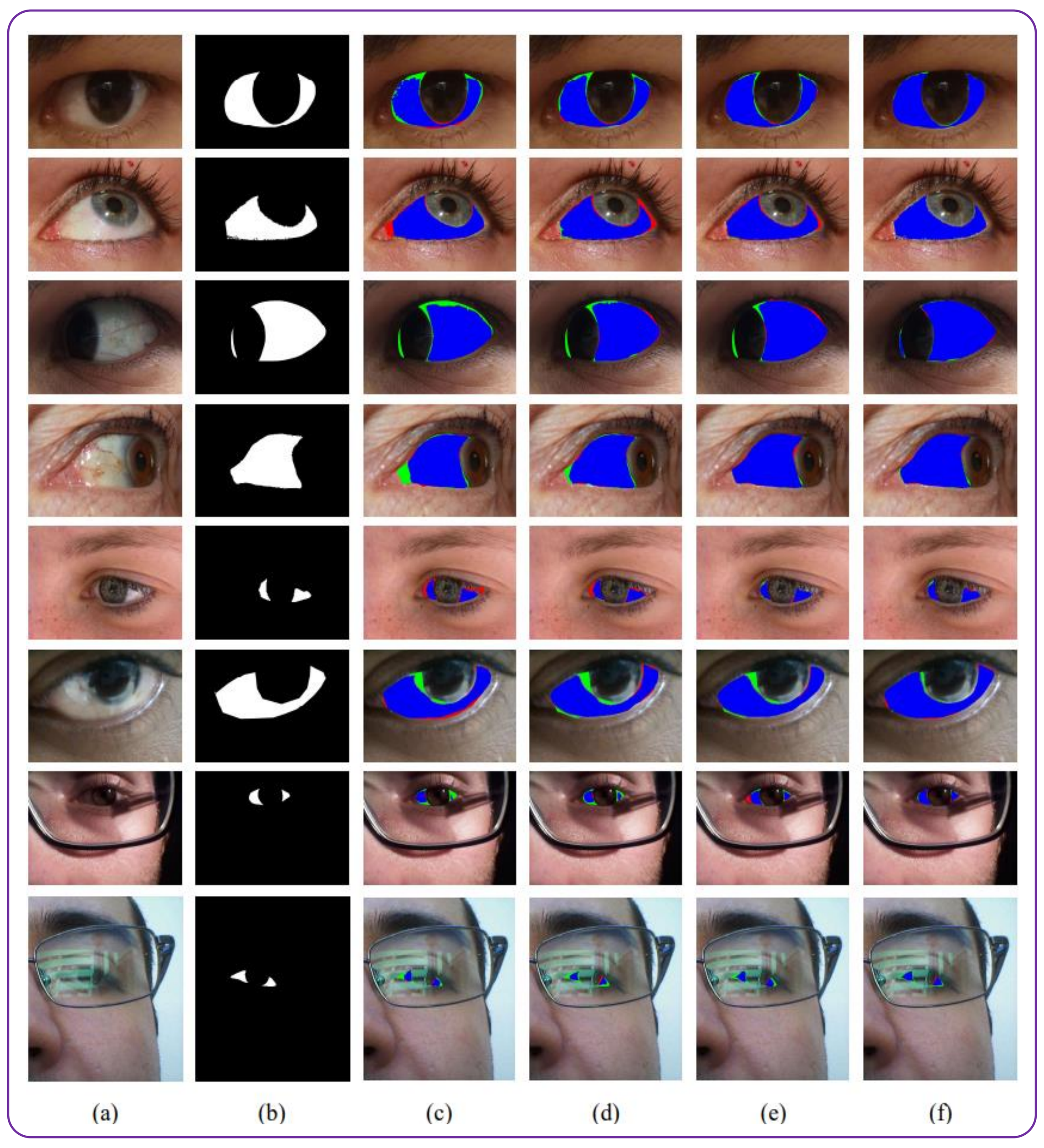
− denotes removing certain block.
△ denotes replacing certain block.

Influence of encoder and CDF

| $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\lambda_4$ | P(%) ↑ | R(%) ↑ | F1(%) ↑ | IOU(%) ↑ |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 96.61(02.12) | 96.79(03.34) | 96.59(02.54) | 93.13(03.51) |
| 0.5 | 0.5 | 0.5 | 1 | 96.59(01.98) | **96.82(03.33)** | **96.66(02.00)** | **93.59(02.85)** |
| 0.1 | 0.1 | 0.1 | 0.7 | 96.50(02.92) | 96.68(04.17) | 96.54(02.63) | 92.89(03.25) |
| 0 | 0 | 0 | 1 | **96.63(02.87)** | 96.54(04.06) | 96.56(02.86) | 92.91(03.11) |

Influence of deep supervision

### Visualized results




Sclera segmentation results of challenging samples.
(a) input eye images, (b) ground truth, (f) segmentation results of our Sclera-TransFuse